Universidad Autónoma de Madrid

Instituto Nicolás Cabrera

Facultad de Ciencias

Desarrollo de una plataforma informática para la búsqueda de nuevos fármacos

Tesis Doctoral presentada por

Álvaro Cortés Cabrera

Director de Tesis

Antonio Jesús Morreale de León

Federico Gago Badenas

Tutor del Departamento

Enrique Velasco

MADRID 2013

Para, y a pesar de, Cristina

AGRADECIMIENTOS

Sin duda esta tesis tiene tres pilares sin los que no hubiera sido lo que es: Antonio, Federico y Cele.

Normalmente, las cosas no se aprecian hasta que desaparecen. Me refiero a la suerte que he tenido de haber conocido a Antonio Morreale. Le debo mucho. Es el mejor de los jefes que podía haber tenido. Su calidad como persona me ha dejado realmente muy impresionado. Antonio no solamente ha tenido que hacerse cargo de un laboratorio en situaciones muy difíciles, sino que, además, nos ha sacado adelante a costa de su tiempo y esfuerzo. Solo espero que se me haya pegado algo.

A Federico le conocí de rebote, aunque sinceramente, me habría gustado haberle conocido en la licenciatura. Sin conocerme, siempre se prestó a apoyarme en mis empeños por llevar a cabo esta tesis y conseguir pasta para ello. Sin duda, este apoyo tiene mucho que ver en como es mi vida ahora, como he podido llevar a cabo proyectos personales que no habrían sido posibles de otra forma, por lo que estoy infinitamente agradecido. Últimamente debo agradecerle también que, cuando la cosa se ha puesto fea, me haya acogido en su laboratorio "*full time*" y me siga apoyando como ha hecho siempre (aunque hoy por hoy podría competir directamente con ACNUR en el tema de los refugiados).

Por último, Cele, ¡qué puedo decir! De Cele no solo he aprendido como científico, sino como persona. Sus consejos me han servido siempre en lo profesional y en lo personal. Todavía me viene la sonrisa cuando pienso en la langosta de Maine con el Nobel ese japonés comiendo enfrente.

No quiero olvidarme tampoco de mis compañeros en el proceso: Pedro, de verdad, ojalá hubiéramos coincidido antes, eres el mejor. Es impresionante el afán de mejorar y hacer cosas que tienes, le pones ilusión a todo y encima se pega. Marta, he de confesar que me gustan más las conversaciones de café cuando vienes por la UAH. Claire, no se me olvida la comida que me debes del Arturo's del CEU.

También son muy responsables de todo este trabajo mis compañeros de la, hoy *zombie*, Unidad de Bioinformática del CBM-SO, Javi y Helena, por las conversaciones existenciales-metodológicas sobre lo que deberían hacer los programas. Alfonso y David, os doy las gracias por haberme aguantado la coña que doy en general y las liadas pardas en particular de petar el cluster regularmente. Rubén y Almudena os agradezco todo los que me habéis enseñado y ayudado. Y al ex-lado oscuro del laboratorio, ahora único lado, Alberto y Raúl por hacer sencillas las respuestas a esas preguntas generales que de vez en cuando os suelto. A Alberto me gustaría agradecerle la oportunidad que nos ha dado en el máster de biofísica y, especialmente, por su comportamiento cuando solicité la estancia en París y pedirle disculpas por todas las movidas que tuvo que aguantar.

A Garrett le escuché un día algo así como "indeed, choosing the right tool for the job is 80% of the work". Supongo que es una de las lecciones más importantes que he aprendido junto con la de que llegaba a las 11.00 al labo y que Mark&Spencer tiene comida preparada.

Aunque su contribución sea más difusa ya que no le conocí, me gustaría agradecer también a Ángel (-= ARO =-). Es increíble que casi 20 años después haya podido sacar tanto partido a su código y métodos, ¡Ojalá algún cacho de código mío pudiera durar

tanto! Muchas veces he acabado escribiendo rutinas y he visto después que él se la curró igual para una cosa similar en su momento (-.-). La verdad que da bastante lástima como se han desarrollado los acontecimientos. Jefes sin grupo y ¿grupo? sin jefe.

A mis suegros y a mi cuñada les doy las gracias por fingir que les parece interesante lo que hago y animarme, sospechosamente, a que abandone el país, en solitario y de una forma regular, y siempre para ir a sitios tristes y fríos o en los que directamente ¿hablan? francés.

Por último (*last but not least*) me gustaría agradecer el apoyo a mi familia. A mis padres y a mi hermana, por poner lo que tengo en la cabeza y por aguantar explicaciones desde siempre de cosas que no les interesan pero fingiendo eficazmente que sí para que yo me lo crea y esté feliz. Es difícil saber que habría pensado mi padre de todo esto, pero nunca me ha gustado perder el tiempo con *ysis* así que asumiré que habría estado contento igualmente y santas pascuas.

Bueno, a Cristina...ya lo sabe ella y además tiene la dedicatoria y media docena de programas pochos con su nombre. ¡El sueño de toda mujer! ¡Qué más quiere!

RESUMEN

El estudio del espacio farmacológico es una tarea muy compleja y la utilización de la gran cantidad de información que contiene con el fin de desarrollar nuevos fármacos obliga a aplicar un gran número de técnicas y aproximaciones

En este trabajo se han aplicado dos rutas diferentes para construir una plataforma computacional que permita el descubrimiento de nuevos fármacos. Por un lado se han desarrollado herramientas para el análisis y la simulación de las interacciones intermoleculares entre los fármacos y sus dianas presentes en los complejos tridimensionales, lo que nos permite predecir la afinidades de unión y proponer nuevas moléculas candidatas.

Por otro lado, se ha introducido un cambio radical en el concepto tradicional de diseño asistido por ordenador al incluir la farmacología de sistemas y el análisis de múltiples variables del espacio químico-biológico para poder estimar perfiles polifarmacológicos, efectos adversos y los complejos caminos de optimización que llevan desde un posible candidato a un nuevo fármaco.

The study of pharmacological space is indeed a complex task because its tortuous nature demands a large number of techniques and approximations to exploit the information that it contains with the aim of developing new drugs.

For dealing with the complexities involved in the construction of a fully functional computational platform for drug discovery In this work we have developed several tools for: (1) simulation and analysis of intermolecular interactions present in three-dimensional complexes comprising compounds and targets; and (2) approaching systems pharmacology and multivariate analysis of chemico-biological space with a view to predicting polypharmacology binding profiles, side effects and complex optimization routes from promising candidates to full drugs.

ÍNDICE

ABREVIATURAS

AMBER Assisted Model Building with Energy Refinement

AUC Area Under the Curve

BEDROC Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic

BEI Binding Efficiency Index

BFGS Broyden-Fletcher-Goldfarb-Shanno

CADD Computer-Aided Drug Design

CATS Chemical Advanced Template Search

COM Center Of Mass

DUD Directory of Useful Decoys

ECFP Extended Connectivity FingerPrints

EF Enrichment Factor

FtsZ Filament Temperature-Sensitive mutant Z

PME Particle Mesh Ewald

GAFF Generalized AMBER Force Field

GARD Generally Applicable Replacement for RMSD

GPU Graphics Processing Unit

GUI Graphical User Interface

HYDE HYdrogen bond and Dehydration

ISM Implicit Solvent Model

LG Levitt-Gerstein

MACCS Molecular ACCess System

MCSA Monte Carlo Simulated Annealing

NHEA Number of HEavy Atoms

NPOL Number of POLar atoms

PDB Protein Data Bank

PSA Polar Surface Area

RMSD Root-Mean-Square Deviation

SEI Surface-binding Efficiency Index

TMScore Template Modeling Score

USR UltraShape Recognition

VS Virtual screening

VSDMIP Virtual Screening Data Management on an Integrated Platform

1. INTRODUCCIÓN

Las interacciones entre biomoléculas (proteínas, ácidos nucleicos, lípidos, metabolitos, etc.) constituyen el lenguaje básico de los sistemas vivos. Numerosas enfermedades están relacionadas con fallos de funcionamiento (Hoshino, Chatani et al. 1999) o ausencia de estas interacciones (Kishnani, Steiner et al. 2006) y, por tanto, estas son de vital importancia. Por otro lado, su estudio permitiría la manipulación de los procesos vitales para tratar de remediar o incluso eliminar los efectos asociados a un determinado proceso patológico (Strebhardt and Ullrich 2008).

Tradicionalmente, esta manipulación se ha llevado a cabo mediante el uso de pequeñas moléculas, que son capaces de interaccionar con las macromoléculas biológicas naturales. En los últimos 30 años, la aplicación de técnicas de bioingeniería ha permitido aumentar el uso de productos de naturaleza proteica, como son los anticuerpos humanizados (Adams and Weiner 2005), las insulinas humanas y otras hormonas de carácter peptídico (Nagle, Berg et al. 2003; Donati and Rappuoli 2013) y las nuevas vacunas.

Sin embargo, las pequeñas moléculas o ligandos siguen siendo de capital importancia en el tratamiento actual de las enfermedades humanas y animales, y ocupan un lugar privilegiado en este aspecto (Hann and Keserü 2012). Su descubrimiento y posterior lanzamiento al mercado es una tarea ardua que requiere una gran inversión en investigación y posterior desarrollo en fase clínica, en promedio, 15 años y 800 millones de dólares por molécula en fase IV (Scannell, Blanckley et al. 2012).

Desde principios de la década de los 90 del siglo pasado, la química combinatoria y el cribado de alto rendimiento han producido un número muy bajo de nuevos fármacos en relación con la inversión realizada. Del análisis de los fármacos aprobados desde 1950, se desprende que el número de éstos ha permanecido prácticamente constante año tras año, indicando que tanto estas tecnologías como la introducción de los compuestos biofarmacéuticos no han supuesto el impacto revolucionario se esperaba.(Munos 2009; Pammolli, Magazzini et al. 2011)

Si bien el grueso de la inversión se debe a los requisitos necesarios para demostrar la seguridad y eficacia de las nuevas moléculas, como son los ensayos clínicos, la fase de investigación pre-clínica es responsable de estos costes añadidos, en cuanto que no impide que candidatos poco prometedores lleguen a fases más tardías de desarrollo, donde el fracaso es más costoso desde el punto de vista económico (Scannell, Blanckley et al. 2012).

El proceso moderno de búsqueda de fármacos comienza con la fase de investigación pre-clínica donde, mediante diversas técnicas, se buscan compuestos que sean capaces de tener un efecto sobre una diana previamente validada por tener un impacto en un proceso patológico. Estos compuestos que interaccionan con una afinidad baja o media se denominan *hits*. Posteriormente, estos compuestos han de ser revalidados mediante experimentos adicionales confirmatorios. Una vez validados, los compuestos entran en un ciclo de optimización-prueba donde se busca conseguir compuestos con mayor afinidad mediante la adición de nuevos grupos químicos a la molécula original. El resultado de este proceso se conoce como *lead* o cabeza de serie. Finalmente, este cabeza de serie entra en nuevos ciclos de optimización donde se busca mejorar su perfil farmacocinético y farmacodinámico, alterando sus propiedades fisicoquímicas y su selectividad por la diana en cuestión frente a otras dianas similares. Finalmente y durante todo el proceso, se realizan pruebas toxicológicas y de seguridad en modelos *in vitro* o en modelos animales, con el fin de intentar extrapolar los resultados al cuerpo humano.

Los métodos teóricos, correctamente entendidos y aplicados, pueden ayudar al químico farmacéutico a diferenciar y elegir los candidatos más prometedores de entre los miles o incluso millones de posibles moléculas disponibles en las quimiotecas a su alcance. Estas técnicas pueden no solo identificar *hits*, sino que también pueden ser empleadas, con ciertas limitaciones, en la optimización de estas moléculas para convertirlas en *leads* y en la consecución de posteriores mejoras que conducirán a los compuestos finales.

1.1 Proceso de unión ligando-receptor

Actualmente hay tres modelos que representan la unión proteína-ligando. El primero de ellos data de finales del siglo XIX y fue postulado por Emil Fischer (Fischer 1894). Se trata del bien conocido modelo de llave-cerradura (*lock-and-key*): solo la llave correcta puede encajar en su cerradura. Se trata de una aproximación muy rígida, ya que no se consideran adaptaciones mutuas entre el ligando y la diana. Una aproximación más flexible, conocida como modelo de acoplamiento inducido (*induced fit*), fue posteriormente propuesta por Daniel Koshland (Koshland Jr 1958). El acoplamiento inducido considera que la flexibilidad intrínseca de la diana se traduce en una reorganización de su centro activo para acomodar a los ligandos entrantes. Por último, el modelo de selección conformacional (*conformational selection*) (Ma, Kumar et al. 1999; Tsai, Kumar et al. 1999) postula que es el ligando el que selecciona, de entre un conjunto de conformaciones accesibles de la diana, la más apropiada para su unión. Una representación pictórica de los tres modelos se puede ver en la figura 1.1.

1.2 Energética de la unión ligando-receptor

La magnitud principal que determina la unión entre un ligando y su diana es la energía libre de unión, que se define como la diferencia de energías libres entre la correspondiente al complejo ligando-diana y la de sus respectivas especies aisladas (ecuación 1.1) con las que se encuentra en equilibrio.

$$\Delta G_{binding} = -RT \ln(K_d) \tag{1.1}$$

siendo *R* la constante de los gases ideales, *T* la temperatura y K_d la constante de equilibrio.

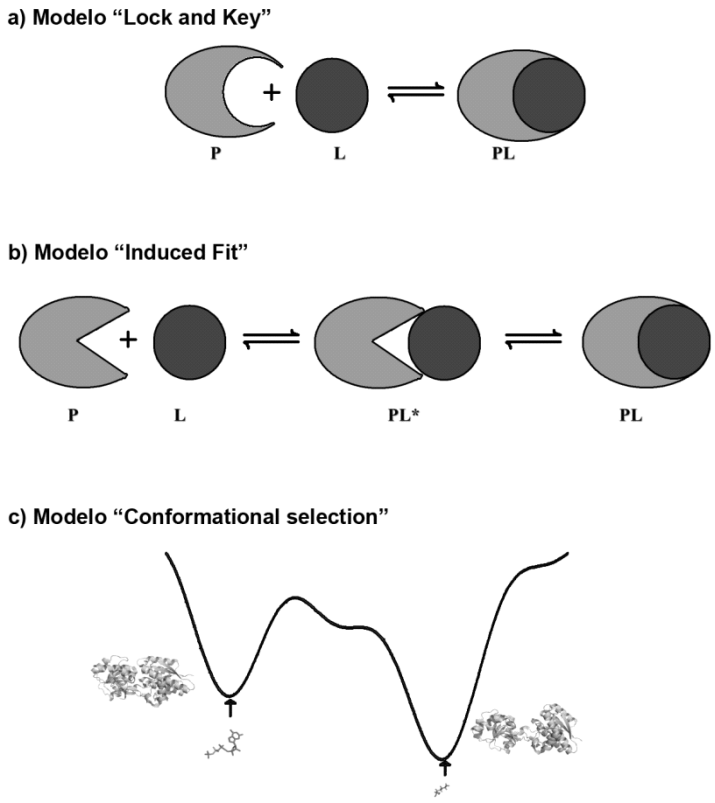


Figura 1.1. Representación gráfica de los tres modelos de unión más comunes.

La energía libre de unión puede descomponerse a su vez en sus componentes entálpica ($\Delta H_{binding}$) y entrópica ($\Delta S_{binding}$).

$$\Delta G_{binding} = \Delta H_{binding} - T\Delta S_{binding} \tag{1.2}$$

La parte entálpica está compuesta fundamentalmente por las interacciones específicas que se producen entre el propio ligando y la diana y entre las moléculas de agua que están presentes en el medio biológico y rodean e interacciones con las dos entidades. Por otro lado, la parte entrópica corresponde al hecho de la pérdida de libertad conformacional que experimentan el ligando y la diana al formar el complejo estable junto con el aumento de libertad que experimentan las moléculas de disolvente que se hallaban interaccionando con ligando y diana.

Los principales tipos de interacciones moleculares que se consideran fundamentales para entender y racionalizar la unión ligando-diana son las siguientes: interacciones de van der Waals (vdW), interacciones electrostáticas, enlaces de hidrógeno, interacciones con el disolvente, interacciones hidrofóbicas y contribuciones entrópicas.

1.2.1 Interacciones de tipo van der Waals (vdW)

Cuando dos moléculas se aproximan y van entrando en contacto las interacciones de vdW dan cuenta dos tipos de fuerzas diferentes: a) repulsión, la cual actúa a corta distancia debido al solapamiento o superposición de las nubes electrónicas de los átomos que se acercan, y b) atracción, que se da a larga distancia por la interacción entre los electrones y núcleos de los diferentes átomos, y se deben más a la forma (o volumen) que propiamente al contenido electrostático. Ambas fuerzas dependen de la distancia entre los átomos (r) por lo que su representación es bastante directa. El modelo más usado es el del potencial de Lennard-Jones, donde el término repulsivo depende de r^{-12} y el atractivo de r^{-6}.

1.2.2 Interacciones electrostáticas

Las interacciones electrostáticas están presentes en la mayor parte de los procesos de unión (interacciones carga-carga, enlaces de hidrógeno, apilamiento de nubes π o π-π stacking, interacciones hidrofóbicas, solvatación, etc.). La aproximación más simple es el modelo Coulómbico (el producto de las cargas dividido por la distancia y una función dieléctrica sencilla simulando el apantallamiento del disolvente).

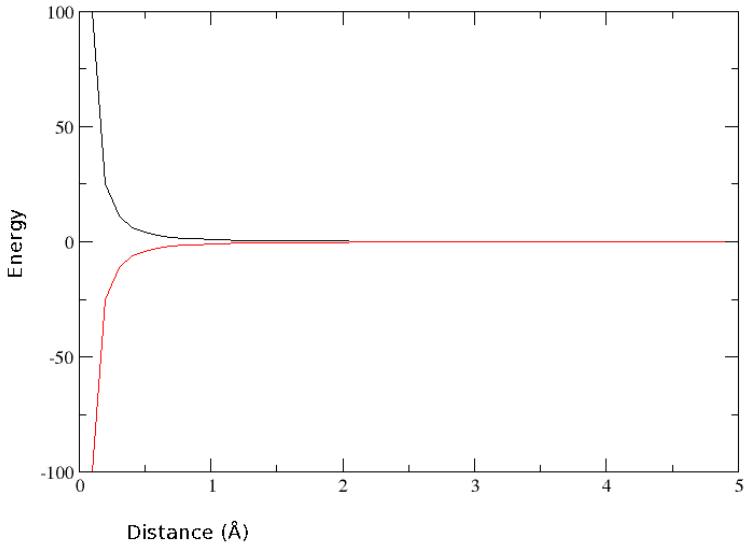


Figura 1.2. Representación del modelo Coulómbico para interacciones electrostáticas entre átomos. En negro, interacción positivo-positivo o negativo-negativo. En rojo, interacción positivo-negativo o viceversa.

1.2.3 Enlace de hidrógeno

Se trata de una interacción muy selectiva y altamente dependiente de la orientación de sus constituyentes. Se establece entre un átomo de hidrógeno unido a un átomo electronegativo, llamado donador de enlace de hidrógeno, y otro átomo también electronegativo, llamado aceptor de enlace de hidrógeno. La fuerza del enlace depende de la posición relativa de los tres átomos implicados, es decir de las distancias y ángulos que haya entre ellos (Liu, Wang et al. 2008). Su papel en el reconocimiento molecular es de gran importancia (Connelly, Aldape et al. 1994).

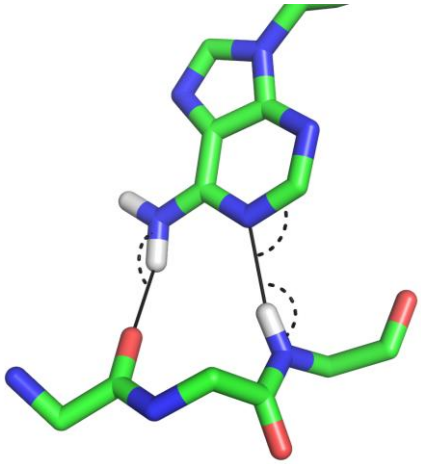


Figura 1.3. Enlaces de hidrógeno y su caracterización geométrica por distancia y ángulos donador-hidrógeno-aceptor e hidrógeno-aceptor-átomo vecino.

1.2.4 El efecto del disolvente

Las interacciones en medio biológico tienen lugar en un entorno acuoso. Cuando las moléculas están aisladas en disolución están completamente rodeadas de moléculas de agua. Sin embargo, cuando se produce la unión ligando-diana muchas de estas moléculas de agua son desplazadas. Este desplazamiento conlleva un gasto energético que debe ser contrarrestado por las nuevas interacciones formadas. Además, se produce una ganancia de entropía en las moléculas de agua liberadas. Desde un punto de vista teórico, hay dos modelos extremos para tener en cuenta los efectos del disolvente: a) modelos de disolvente explícito, donde las moléculas de agua están explícitamente representadas a detalle atómico, y b) modelos de disolvente implícito, donde se construye una función matemática que trata de simular el comportamiento global del disolvente. También es posible considerar modelos mixtos en los cuales se tienen en cuenta explícitamente determinadas moléculas y el resto se consideran de manera implícita.

1.2.5 Interacciones hidrófobas

Determinados constituyentes de las dianas, como son las cadenas laterales de algunos aminoácidos (leucina, valina, prolina…) al igual que muchos ligandos están formados por grupos o estructuras químicas de naturaleza apolar, que interactúan de manera muy pobre o desfavorable con moléculas polares como el agua en el que están inmersas. En el evento de la unión, si dos superficies hidrófobas interaccionan, dan lugar a una liberación de las moléculas de agua de ambas superficies, lo que contribuye a incrementar la entropía del sistema favoreciendo la unión.

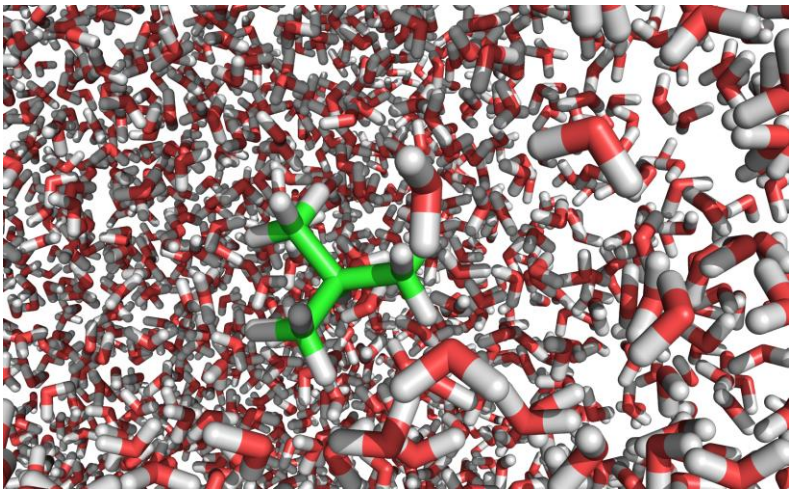


Figura 1.4. Simulación con disolvente explícito. El ligando apolar (C, verde; H, blanco), se encuentra en un entorno desfavorable de moléculas de agua (O, rojo; H, blanco).

1.2.6 Contribución entrópica

La formación de un complejo diana-ligando lleva asociada una pérdida de libertad de movimiento por parte de ambas entidades que se traduce en una disminución de la entropía. Aparte de la entropía del disolvente (que incrementa al liberarse del ligando y la cavidad de la diana), la entropía del soluto (o entropía configuracional) se suele dividir en dos partes: conformacional y vibracional. La parte conformacional tiene que ver con la reducción del número de pozos de energía que tanto el ligando como la proteína pueden visitar una vez que ha sucedido la unión, mientras que la parte vibracional se refiere a los movimientos dentro de un pozo de energía en particular. La estimación de la entropía es muy compleja, y los márgenes de error suelen ser sensiblemente más altos que los de otros términos (Hou, Wang et al. 2010).

1.2.7 Otras interacciones

Existen otras interacciones que pueden ser tratadas de manera explícita según el modelo utilizado. Estas pueden incluir enlaces de hidrógeno no convencionales (-CH$^{\cdots}$ aceptor), enlaces con halógenos (Politzer, Murray et al. 2010), etc., los cuales pueden también ser contemplados implícitamente en el resto de términos tratados con anterioridad.

1.3 Farmacología de sistemas

Esta disciplina ha emergido como consecuencia de la aplicación de los principios de la biología de sistemas al campo de la farmacología (Zhao and Iyengar 2012). Se fundamenta en la aplicación de aproximaciones teóricas, como el análisis de redes a múltiples escalas de organización biológica (Berger and Iyengar 2009) (Figura 1.5), y experimentales, como son los datos disponibles de genómica y proteómica en las grandes bases de datos, para dar una visión global de la acción de los principios activos en un contexto general que incluye desde unas pocas vías de señalización a todo un genoma (Arrell and Terzic 2010).

La farmacología de sistemas ha facilitado el desarrollo del concepto *drug repurposing* o reutilización de principios activos ya conocidos (Oprea, Bauman et al. 2012). También ha permitido avanzar en el estudio de las reacciones adversas de medicamentos, al considerar en su conjunto la red de posibles dianas que existen en el organismo y que pueden interaccionar con una molécula dada (Lounkine, Keiser et al. 2012).

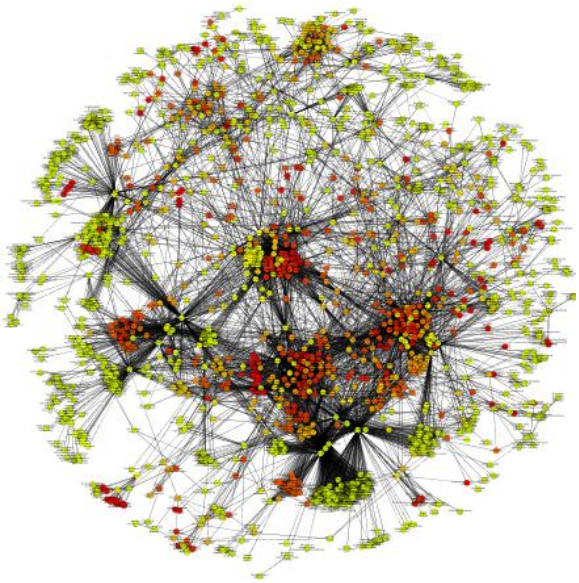


Figura 1.5. Red compleja que resulta de analizar un conjunto de interacciones farmacológicas

1.4 El espacio químico-biológico (*Chemical-Biological Space*, CBS)

Los principios activos y las dianas sobre las que actúan conforman un espacio en el que su relación (afinidad) y sus propiedades fisicoquímicas (grupos químicos, polaridad, masa, permeabilidad a barreras biológicas, etc.) son las principales variables que lo definen.

Este espacio químico (por parte de los ligandos o principios activos) y biológico (dianas en el paciente) se explora continuamente en el proceso de desarrollo de fármacos mediante la síntesis o descubrimiento de nuevas moléculas y su evaluación biológica posterior sobre la diana correspondiente. Sin embargo, la gran complejidad del CBS hace muy difícil la optimización de compuestos, por hoy día sigue siendo un proceso costoso y dirigido mediante técnicas de ensayo-error (Scannell, Blanckley et al. 2012).

Su exploración sistemática bajo un marco bien definido de variables es lo que se conoce como el AtlasCBS (Abad-Zapatero, Perisic et al. 2010), en el que determinados subespacios del CBS (dianas, tipos de compuestos, etc.) son agrupados y mostrados en planos, los cuales permiten una rápida evaluación de sus propiedades químicas y biológicas.

2. MATERIALES Y MÉTODOS GENERALES

2.1 Classical molecular mechanics

Biological systems of interest usually are made up of millions of atoms, and even focusing on small parts the number of particles to deal with exceeds the tens of thousands. In fact, despite the continued growth in computational power, molecular modellers still have to rely on convenient approximations reducing the complexity of the systems under study (Barril and Luque 2012). In this context, the key components of the atoms are decoupled and simplified. According to the Born-Oppenheimer approximation (Born and Oppenheimer 1927), the electrons move faster than the core of nuclei, as a consequence of their reduced mass. Therefore, from the point of view of the electrons, the nuclei are in fact frozen, and from the point of view of the nucleus, the electrons act like a cloud that responds immediately to a change in the coordinates of the core. Relaying on the difference in masses it looks reasonable to decouple both components and to describe the atoms as spheres neglecting the need to consider the electrons and their complex interactions explicitly.

The downside of this approximation is that the vast majority of the interesting problems at the molecular level appear as a result of the behaviour and interactions of the electrons. For this reason extra terms and *tricks* have to be included in order to reproduce these effects to some extent at least.

2.1.1 Force Fields

The forces generated due to the interactions among atoms can be estimated through an empirical potential function based on a group of parameters from *ab initio* calculations, experimental results or a combination of both. This combination of potential and parameters is known as the force field. Several force fields have been developed since the 70s (Mayo, Olafson et al. 1990; Cornell, Cieplak et al. 1995; Halgren 1996) and, in general, they introduce the potential as a sum of the bonding and non-bonding terms. Bonding terms try to reproduce the effect of the covalent bonds on the geometry of the molecule while the non-bonding terms try to reproduce the interaction between not bonded parts of a molecule or between different molecules.

Bonding interactions, although depending on the force field, include three different terms: the stretching, the angle and the dihedral potentials.

The stretching potential simulates the covalent bond between two spheres allowing a stretching movement between them:

$$E_{stretching} = \sum_{i}^{\#bonds} \frac{1}{2} k_i \left(r_i - r_{0i}\right)^2 \quad (2.1)$$

where k_i is a constant that represents the strength of the atomic bond, r_i the distance between the atomic centers and r_{0i} the ideal distance for that bond.

The angular restrictions imposed by the atomic orbitals are taken into account with the angle potential:

$$E_{angle} = \sum_{i}^{\#angles} \frac{1}{2} k_i (\gamma_i - \gamma_{0i}) \quad (2.2)$$

where k_i is a constant that represent the flexibility of the angle, γ_i the actual angle between the three atomic centers and γ_{0i} the ideal angle for those three atoms.

Finally, rotation about single bonds usually give rise to several minima and maxima in the potential energy function. To represent this phenomenon, the harmonic model is not suitable and the most popular approach is to represent it by means of a Fourier series:

$$E_{dihe} = \sum_{i}^{\#dihedrals} \frac{1}{2} V_n (1 + \cos(n\varphi - \gamma_n)) \quad (2.3)$$

where n is the current minimum of the potential (commonly with a maximum of 3 or 4), V_n the Fourier coefficients, φ the dihedral angle of the torsional and γ the phase.

In addition, it is common to add improper dihedrals terms to correct out-of-plane deviations (e.g. benzene ring).

Non-bonding term is computed over all non-bonded atom pairs (except for 1-3 atoms and 1-4 atoms where the potential is divided by 2) and it is decomposed into two different terms: van der Waals (12-6 potential calculated with a equilibrium distance and a well-depth that depends on the pair of atoms) and the electrostatic contribution according to Coulomb's law.

$$E_{nonBonding} = \sum_{i}^{Npairs} \left[\frac{A_i}{r_i^{12}} - \frac{B_i}{r_i^{6}} + \frac{q_{1i} q_{2i}}{\varepsilon r_i} \right] \quad (2.4)$$

where i is the pair of atoms, A_i and B_i are the force field coefficients for that pair of atoms, r_i the distance between the atoms, q_{1i} and q_{2i} the charges of the atoms and ε the dielectric function or constant.

2.1.2 Energy minimization

Using the force field potential and function minimization algorithms such as conjugate gradients (Hestenes and Stiefel 1952) or steepest descent (Petrova and Solov'ev 1997) it is possible to optimize the geometry of a molecular system to be compatible with the parameters of the force field and to reach a minimum of energy in the potential surface. It is a necessary step prior to computationally more demanding tasks such as molecular dynamics or docking.

2.1.3 Molecular dynamics

Molecular dynamics simulations start from an initial spatial configuration of the system of interest (a theoretical or experimental model) and new states are generated using the Newton's equations of motion for certain time steps.

Introduced for the first time at the end of the 50s (Alder and Wainwright 1959), (Alder and Wainwright 1959), its popularization and extensive use did not occur until the advent of the digital era due to its demanding computational cost. It is noteworthy that nowadays molecular dynamics is one of the most employed simulation tools that is performed on supercomputers all over the world. Furthermore some specific computation machines have been developed to solely perform this task (Shaw, Deneroff et al. 2007).

Once the force field potential is defined, Newton's equations are applied:

$$\begin{cases} \dfrac{p_i}{m_i} = \dfrac{dr_i}{dt} \\ \dfrac{dp_i}{dt} = F_i \end{cases} \tag{2.5}$$

where p_i, m_i, r_i and F_i are the linear moment, the mass, the position and the force over the atom i, respectively.

A system with N atoms will presents $6N$ first order differential equations ($3N$ coordinates) that can be solved by using the finite differences method to obtain a trajectory of each atom of the system. The integration time step is crucial and it is often reduced to the highest vibration frequency of the system, 0.5 fs. However, it is still possible to reach 2 fs if an algorithm to freeze the atoms responsible of these high frequency vibrations is used (Andersen 1983; Hess 2008). This reduced time step is several orders of magnitude smaller than the time scale of most interesting molecular phenomena such conformational changes, protein folding or ligand binding (Pande, Baker et al. 2003).

The most time consuming part of the simulation is the evaluation of the non-bonded terms and the update of the list of pairs. For this reason, the algorithms employed try to avoid the update of the list at each step and use non-*naive* algorithms that limit the candidate atoms to those which are inside a sphere where the potential to evaluate is not null in practice (van der Waals) or switch to an alternative faster algorithms at a given distance [Particle Mesh Ewald for long-range electrostatic interactions (Darden, York et al. 1993)]. In addition, with the advent of the Graphics Processors Units (GPU), massive multi-parallel calculations on those small devices with optimized kernels have increased the simulations lengths several orders of magnitude making it possible to reach the time scale of microseconds on a regular basis and often milliseconds (Götz, Williamson et al. 2012).

2.2 Docking

Given the three dimensional structure of a target of interest, we could define molecular docking or just docking as the search of the most suitable 3D complex of a ligand and the target.

The ligand is usually able to adopt several spatial configurations (or poses) inside the binding pocket of the macromolecule (usually a protein). For this reason the docking program should be designed to explore these possibilities (search space) and to evaluate the goodness of these poses by taking into account the propensity of similar interactions in nature (scoring function).

The docking program could be reduced to an algorithm that navigates efficiently through the search space composed of all possible placements of the ligand within a 3D space and evaluate each of these poses according to the observed experimental interactions and the chemical nature of the species involved in these interactions.

Depending on the definition of the search space, the number of combinations to try could exceed easily the computational power that is available today. For this reason, any simplification of this space could be of great importance to enhance the accuracy of the tools and to improve the execution times. Usually, the search is restricted to a region of the space, the binding pocket, where the modeller estimates that the interaction is more likely to take place due to additional information available on the target (enzyme, catalytic residues, other binders, etc.). However, a full search for the whole target is still possible when the user has no information regarding the binding site. In this latter case blind docking is performed.

The degrees of freedom considered in the binding event also have a severe impact on the speed and reliability of the process. If the two entities involved in the interaction are

considered as two rigid bodies (rigid docking), the search space is then reduced to six dimensions (three for translation and three for rotation of the ligand, which is the smallest molecule). If we also consider internal flexibility of the ligand and/or the internal flexibility of the macromolecule we are talking about fully flexible or semi-flexible docking, respectively.

2.2.1 Rigid docking

The simplification of considering the entities to two rigid bodies has its roots in the key-and-lock interaction paradigm (Fischer 1894). The main advantage of this procedure is the large reduction of the search space (six dimensions) that even allows a brute force (exhaustive) algorithm to be run in a few seconds. The main problem arises when the ligand conformation is considered. Usually, relaxed or low-energy conformations are used, but in some cases there are several possibilities, or in the worst case scenario, the conformation of the best interacting complex is in a high-energy state imposed by the target. For this reason, rigid docking is not very popular nowadays although it was the only docking option available at the beginning of the CADD field (Kuntz, Blaney et al. 1982). However, it is still in use to validate new tools for docking and it will be used in the work described below.

2.2.2 Semi-flexible docking

Small molecules, except for the simplest rigid entities, possess some degree of internal flexibility. This is mainly due to the fact that low rotational barriers about their torsional angles allow them to rotate at the typical temperatures at which binding takes place. According to the choice of the algorithms to simulate this behavior they can be divided into multi-rigid and truly flexible approaches. The former use a rigid-docking engine with multiple conformers generated previously according to different methods (Monte Carlo Simulated annealing, exhaustive enumeration, rule-based, etc.). They tend to be faster and very scalable since the docking is reduced to a number of rigid dockings (McGann 2012). Truly-flexible algorithms explore the torsional degrees of freedom of the ligand in the search process adding an extra complexity to the problem and to the algorithms (Morris, Huey et al. 2009). Thus, the main advantage of this method is that if the best ligand pose is not present in an appreciable percentage of the ligand conformations in solution, the algorithm is still able to generate it and find the proper solution at the cost of lower performance.

2.2.3 Fully flexible docking

In fully flexible docking the internal degrees of freedom of the protein are also considered. These approaches can be divided into two classes depending on the degree of motion that is being simulated. Some targets only carry out a limited rearrangement of their binding

pocket side chains to accommodate ligands during/after the binding event. In this case, exploring only those side chains known to be flexible could improve the runtimes and provide a proper approximation for these targets (Sherman, Day et al. 2006). In other cases, targets experience a large rearrangement, involving conformational changes (e.g. hinge movements) (Wang and Merz Jr 2010) that cannot be approximated with the limited flexibility of the side chains. In these cases, protein flexibility is explored prior to docking and generating a series of conformations, typically from molecular dynamics simulations or normal mode analysis (Totrov and Abagyan 2008). These conformations are then used in a multi rigid protein-docking fashion.

2.3 Search algorithms

Throughout the history of docking, several algorithms have been applied to this task and new ones are constantly being developed and introduced (Kuntz, Blaney et al. 1982; Chen, Liu et al. 2007; Morris, Huey et al. 2009; Cabrera, Klett et al. 2012). In this section only those that were used in our own work will be described.

2.3.1 Discretization of the space

As a first step in the search algorithms some of them require a discretization of the search space. In this process the binding pocket, or the area selected for the search, is evaluated according to a scoring function at certain distances in the three dimensions (usual values are 0.25 Å, 0.375 Å, 0.5 Å or 1.0 Å, depending on the resolution requested), generating the energy, potential or scoring function grids that are stored for further use. In the program runtime, these values are loaded into memory and used for, typically, trilineal interpolation to generate scoring values for those positions not located at grid points.

These grids speed up the docking process several orders of magnitude due to the avoidance of full scoring function calculations at the cost of small errors in the scoring values. Nevertheless, it imposes an important constraint on the docking process. Since pre-generated values are used, no flexibility is allowed for the target unless some parts of the macromolecule are sampled together with the ligand, increasing the computational costs.

Some algorithms that sample in real space are also forced to discretize the rotational space. In this manner, the tools can generate, in advance, a set of rotation matrices to be applied to the ligand in a given manner (randomly, exhaustively, etc.), reducing again the search space to a limited number of operations that could be performed in a reasonable amount of time.

2.3.2 Exhaustive search

It is the slowest method but the simplest to implement. The algorithm discretizes the translations, or takes advance of previous discretization at the grids, and performs trials translating the center of mass (COM) of the ligand to the designated place. Simultaneously, the program samples the rotational degrees of freedom (Euler angles) increasing their values in a fixed amount of degrees each time or using quaternions.

Torsional flexibility could also be simulated if multiple ligands are used, repeating the algorithm for each input conformation and storing the results in a common pool.

Exhaustive search

```
1: procedure ExhaustiveSearch(a, b)        ▷ Perform an exhaustive search
2:     for i ← 1, Xpoints do
3:         for j ← 1, Ypoints do
4:             for k ← 1, Zpoints do
5:                 destination ← 0, 0, 0
6:                 destination ← TranslateToXYZ(i, j, k, gridCoordinates, spacing)
7:                 Remove COM of LIG                    ▷ Center LIG at 0,0,0
8:                 for r ← 1, Rotations do
9:                     ApplyRotation RotationMatrices[r] to LIG
10:                end for
11:                Move COM of LIG to destination
12:                Energy ← EvalScoreofLIG
13:                if Energy is less than LastMinEnergy or Stack not full then
14:                    KeepPose
15:                else
16:                    RejectPose
17:                end if
18:            end for
19:        end for
20:    end for
21: end procedure
```

Algorithm 2.1. Docking Exhaustive Search

2.3.3 Triangle search

This algorithm was developed with the aim of reducing the search space to provide a faster tool without losing possible solutions. It is based on the idea that not all of the positions in the search space would be compatible with the chemical features of the ligand, e.g. hydrogen bond acceptors are expected to be near hydrogen bond donors establishing a hydrogen bond and not near other hydrogen bond acceptors that could result in a repulsive electronic interaction. In a first step, the algorithm needs to map the designated search space using three different probes: carbonyl (C=O), to discover spots where a hydrogen bond acceptor would establish a favorable interaction, amine (N-H), to find suitable for a hydrogen bond donor, and methane (CH_4) for any other areas with a hydrophobic nature. Spots are filtered according to a simple scoring function and only the best non-redundant probes are kept as possible placement sites of the ligand chemical features.

After the initial analysis of the pocket, the algorithm detects polar features at the ligand and creates a list of triangles with all these points that fulfill a minimum edge length. The triangle has to be able to superimpose the ligand with three different points at the binding pocket, also forming a triangle with the required conditions. Each triangle of the ligand is superimposed onto each triangle at the binding pocket. This algorithm has the advantage of being faster than a traditional exhaustive search and also avoids the discretization of the rotations that could miss some of the solution. The main problems are: i) the strong dependence on the binding zone analysis that could miss an important spot and ii) it is not applicable when the ligand has less than three different polar features (nitrogen or oxygen atoms). In this latter case, an alternative algorithm is required.

Triangle search

```
1: procedure TRIANGLESEARCH(a, b)        ▷ Perform a triangle based search
2:     for i ← 1, LigandPolarAtoms do
3:         for j ← i + 1, LigandPolarAtoms do
4:             for k ← j + 1, LigandPolarAtoms do
5:                 LengthIJ ← √(XYZ[i] − XYZ[j])²
6:                 LengthIK ← √(XYZ[i] − XYZ[k])²
7:                 LengthJK ← √(XYZ[j] − XYZ[k])²
8:                 if LengthIJ < MinLength or LengthIK < MinLength or
   LengthJK < MinLength then
9:                     Continue loop
10:                else
11:                    Add Triangle i,j,k to the triangle stack
12:                end if
13:            end for
14:        end for
15:    end for
16:    for i ← 1, DetectedSpots do
17:        for j ← i + 1, DetectedSpots do
18:            for k ← j + 1, DetectedSpots do
19:                LengthIJ ← √(XYZ[i] − XYZ[j])²
20:                LengthIK ← √(XYZ[i] − XYZ[k])²
21:                LengthJK ← √(XYZ[j] − XYZ[k])²
22:                if LengthIJ < MinLength or LengthIK < MinLength or
   LengthJK < MinLength then
23:                    Continue loop
24:                else
25:                    Add Triangle i,j,k to the other triangle stack
26:                end if
27:            end for
28:        end for
29:    end for
30:    for i ← 1, LigandTriangles do
31:        for i ← 1, TargetTriangles do
32:            SuperImposeTriangle i to j
33:            Energy ← EvalScoreofLIG
34:            if Energy is less than LastMinEnergy or Stack not full then
35:                KeepPose
36:            else
37:                RejectPose
38:            end if
39:        end for
40:    end for
41: end procedure
```

Algorithm 2.2. Docking Triangle Search

2.3.4 Monte Carlo Simulated Annealing (MCSA)

MCSA belongs to the family of stochastic optimization algorithms (Vanderbilt and Louie 1984). It could be applied to those cases where the exhaustive search is not possible due to lack of resources or time constraints.

The algorithm generates random transformation, or a set of random values within the search space, *i.e*. a translation, a rotation and a conformer (only in the case of multiconformer docking to simulate ligand flexibility) used as the starting pose. Each round of the algorithm is limited to a maximum number of steps or to a logical condition (*if this is met first*), *e.g*. No new better pose in 10 trials. At the end of each round, the *temperature* is scaled down by a certain value predetermined for optimum performance. Finally, at each step within a cycle, parameters for the best conformation are transformed by a random value to generate a new pose only if the parameters are set for transformation according to a new random number. The pose is evaluated according to the Metropolis criterion that accepts all new poses with a score better than the rest of the poses in the stack and also, to avoid local minima traps, it accepts high energy poses with a certain probability depending on the temperature of the current cycle. The temperature is high at the beginning to better explore the space and it ends low to restrict the search at the bottom of the global minimum (if found).

MonteCarloSimulatedAnnealing search

```
1: procedure MCSASEARCH(a, b)        ▷ Perform a Monte Carlo Simulated
   Annealing based search
2:     temp ← 800
3:     for i ← 1, nCycles do
4:         for j ← 1, MaxSteps do
5:             for k ← 1, nVariables do
6:                 Rnd ← GenRandom(0 − 1)
7:                 if Rnd < 0.8 then
8:                     LigandCurrentVar ← LigandCurrentVar + (GenRandom() mod (2 − −1) − 1)
9:                 end if
10:            end for
11:            Energy ← EvalScoreofLIG
12:            if Energy < BestEnergy then
13:                Keep pose
14:            else
15:                δE = BestEnergy − Energy
16:                if −δE/temp < GenRandom(0 − 1) then
17:                    Keep pose
18:                end if
19:            end if
20:            if NotNewMinimum in 10 steps then Break ▷ Go to next cycle
21:            end if
22:        end for
23:        temp ← temp ∗ scaling
24:    end for
25: end procedure
```

Algorithm 2.3. Monte Carlo Simulated Annealing Search

2.4 Optimization algorithms

It can be argued that all the algorithms described above could also be classified as optimization algorithms since docking is nothing but an optimization process of a given scoring function. In this section, for the sake of convenience, these optimization algorithms only refers to those that are usually employed in the last steps of the process (local optimizers) while the search algorithms involve global optimizers that try to navigate efficiently in a vast search space.

2.4.1 Nelder–Mead or downhill simplex method

Also known as the *amoeba* method, it was first introduced by Nelder and Mead in 1965 (Nelder and Mead 1965) and aims for optimization of function in a *N*-dimensional space.

The method starts initializing *N*+1 different random variation over the initial pose, where *N* is the number of variables (e.g. translation vector, rotation angles or quaternions and torsionals). Then it ranks the new poses according to their score and selects the two worst of them. If the difference in energy between the best and the worst poses is less than a cutoff, the algorithm ends. If not, it computes the average vector of all simplex points excluding the worst point and performs the ***reflection*** step (multiplied by a factor α) over the worst pose. If the new pose energy is better than the best energy so far, the worst point is replaced and a new reflection, called ***expansion***, multiplied by a larger factor (γ) is attempted. If successful, the new point replaces the old one; if not, the algorithm continues.

If the reflection failed to find a new best point, but the new pose is better than, at least, any of the current members of the simplex, the worst pose is replaced with the new one. If the point is worse than the worst so far, a ***contraction*** step is attempted and if it works, the new pose replaces the old one. Finally if all other steps fail, ***reduction*** is carried out by means of which the poses in the simplex are averaged out with the best one.

The main advantage of this method is that unlike steepest descent or conjugate gradients it does not require the computation of the gradient, which can be costly or even not possible. On the contrary, it usually requires more scoring function evaluations than other more modern methods and this makes it significantly slower.

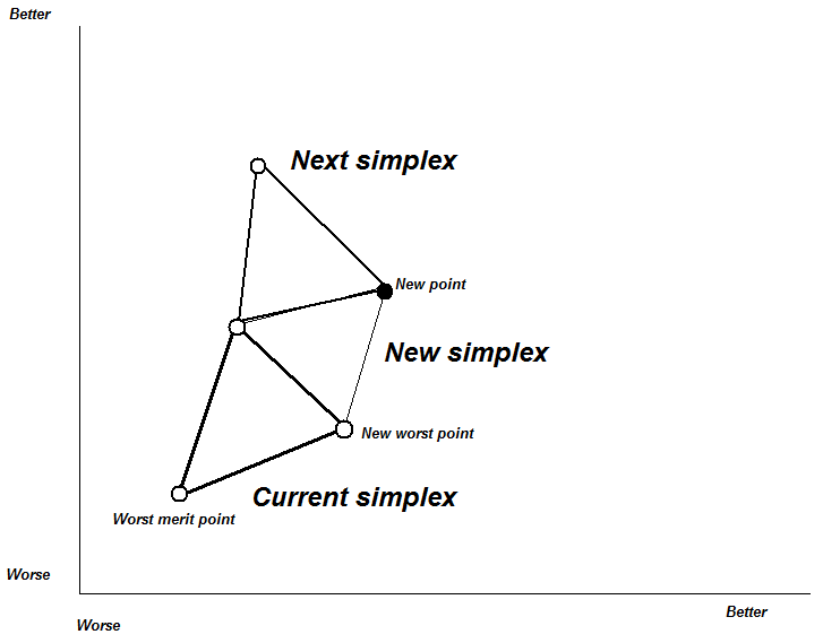


Figure 2.1. 2D simplex simulation. Adapted from Wang et al.(Wang and Shoup 2011)

2.4.2 Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm

It is a quasi-Newton iterative algorithm for nonlinear problems (Head and Zerner 1985). It uses the first and the second derivatives but the latter do not need to be evaluated and instead they are approximated with increasing accuracy.

The algorithm departs from the initial pose to be optimized and calculates the gradient of the scoring function to find the first search direction. Then, it performs a linear search in the opposite direction of the gradient for an appropriate step size, and moves the pose to a new and better point. In the following steps, the search direction is recalculated using the first derivative and the approximated Hessian matrix. At the end of each step, the Hessian matrix is updated with the information available at the current and at the last step gradients.

This algorithm converges faster than the simplex method but it is not able to obtain good solutions if the initial pose is not close to the minimum (Fletcher 1980).

2.5 Scoring functions

Scoring functions are mathematical functions that try to approximate or simulate the behavior and the molecular interactions that occur between ligands and targets in the binding event. Most of them are also designed to be efficient to compute and evaluate and tend to be differentiable (or piecewise differentiable) so that they can be used with fast optimization algorithms.

According to the data used to derive the function and parameters, scoring functions are classified into three classical groups: force field-based functions, empirical functions and knowledge-based functions.

2.5.1 Force field-based functions

The idea behind these functions is the use of the equations and parameters derived for the force fields (see classical molecular mechanics section) to evaluate the goodness of the interaction between ligand and target atoms. Only a reduced form of the force field equation is used, usually the non-bonding interaction part of it plus the dihedral terms. The former tries to capture the effects of attraction and repulsion between atoms, whereas the latter describes the resulting strain energy of a ligand-target interaction in order to avoid those poses with unusual torsional angles and/or with a low probability to exist.

In the case of the present work, the Generalized AMBER force field (Wang, Wolf et al. 2004) non-bonding terms were used:

$$\Delta G = \sum_{i}^{N}\sum_{j}^{n}\left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}}\right] \qquad (2.6)$$

where N is the number of atoms of the target involved in the interaction; n the number of ligand atoms; A_{ij} and B_{ij} the van der Waals parameters defined by the force field which depend on the atoms types of i and j; and q_i and q_j the partial atomic charges of atoms i and j, respectively.

These functions require the typing and evaluation of every singe atom to identify and assign the proper parameters in order to evaluate the function. In some cases, this requirement forces to *perceive* the molecule, *i.e.* to evaluate the connectivity and the chemical nature of the atoms involved in the interaction. The main advantages of these functions are that each term has a physical meaning and that their evaluation can lead to important information for binding pocket analysis.

2.5.2 Empirical functions

These functions are based on terms that try to account for different parts of the interaction that can be adjusted and parameterized using binding data and multiple linear regression analysis. These terms, are then balanced according to a training set of receptor-ligand complexes and tend to represent meaningful and well-known interactions such as hydrogen bonds, non-polar contacts or ionic bonds.

The ChemScore function was proposed by Eldridge and colleagues in 1997 (Eldridge, Murray et al. 1997) for the prediction of free energies of binding. Later, it was improved by Verdonk et al in their docking program GOLD (Verdonk, Cole et al. 2003) with the following functional form:

$$\begin{aligned}\Delta G &= \Delta G_0 + \\ &\Delta G_{hbond} \sum_{i}^{N} g_1(\Delta r) g_2(\Delta \alpha) g_3(\Delta \beta) + \\ &\Delta G_{metal} \sum_{j}^{M} f_m(\Delta r_{jM}) + \\ &\Delta G_{lipo} \sum_{k}^{L} f_l(\Delta r_{kL}) + \\ &\Delta G_{rot} H_{rot}\end{aligned} \tag{2.7}$$

where ΔG_0 is a constant, ΔG_{hbond}, ΔG_{metal}, ΔG_{lipo},and ΔG_{rot} the corresponding regression coefficients and H_{rot} the number of rotatable bonds in the ligand. This last term tries to account for the entropic penalty that the binding event imposes over the freely rotatable bonds by limiting their rotational freedom. The first summation runs for all the possible hydrogen bond pairs. It calculates a score between 0 to 1 on the basis of block functions for the distance and the donor-hydrogen-acceptor angle and hydrogen-acceptor-any neighbor atom angle. These three functions try to characterize the strength of the hydrogen bond as these interactions are primarily directional. The second summation runs for all the acceptor atoms in the ligand and all metal atoms in the binding site, and it only depends on the distance between them. Improved versions of this function include an angle correction for certain metal atoms to reproduce their spatial configuration as found in crystals. Finally, the last summation runs over all pairs of lipophilic atoms and only depends on the distance.

Block functions are characterized by two parameters and have the following functional form:

$$F(x, x_{ideal}, x_{\max}) = \begin{Bmatrix} 1 \quad if \quad x \le x_{ideal} \\ 1 - \dfrac{x - x_{ideal}}{x_{\max} - x_{ideal}} \quad if \quad x_{ideal} \le x \le x_{\max} \\ 0 \quad if \quad x > x_{\max} \end{Bmatrix} \tag{2.8}$$

where x_{ideal} is the ideal value for the parameter and x_{max} the maximum allowed value for the parameter to consider that an interaction exists. This function returns values between 0, no interaction at all, to 1 when the ideal values for the interaction are met.

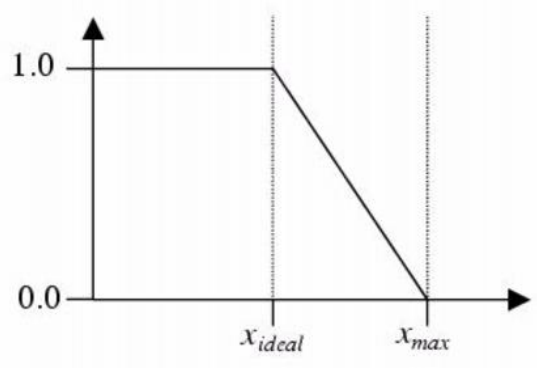


Figure 2.2. Block function representation. Adapted from GOLD docking program documentation(Verdonk, Cole et al. 2003).

CRScore (Cabrera, Klett et al. 2012) is a simplified version of the scoring function GlideScore (Friesner, Banks et al. 2004) which was derived from ChemScore and empirical fitting of some force field terms. It scales down the GAFF van der Waals (E_{vdw}) and Coulomb terms (E_{qq}) by two factors and sums directly both ChemScore hydrogen bond and lipophilic terms.

$$\Delta G = 0.065E_{vdw} + 0.130E_{qq} + \Delta G_{hbond} + \Delta G_{lipo} \tag{2.9}$$

The van der Waals term rises sharply as the distance between the atoms falls below the equilibrium distance and becomes strongly repulsive. By scaling down this term we avoid the overpenalization of near-optimum solutions in which one or more some ligand atoms are close to one or more target atom. The Coulomb term does not rise as fast as the vdW term but, as it tends to dominate the interaction with strong negative or positive values, it also needs to be scaled down by a factor. Finally, the hydrogen bond term rewards poses that present optimal hydrogen bondsings interactions according to their geometry while the lipophilic term accounts for the desolvation of non-polar surfaces of both target and ligand.

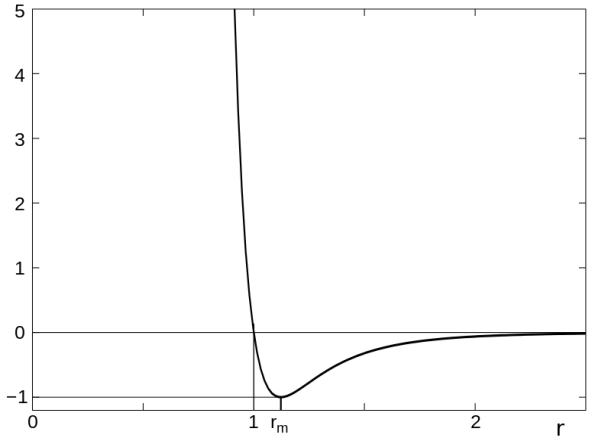


Figure 2.3. Lennard-Jones potential

2.5.3 Statistical potentials

Statistical potentials are based on distributions of intermolecular distances between chemical features or atoms derived from large three dimensional databases, usually the Protein Data Bank (Berman, Westbrook et al. 2000) or the Cambridge Structural Database (Allen 2002). These potentials exploit the information available in crystals of ligands and/or ligand-target complexes and convert it into distance-dependent Hemholtz free interaction energies of protein-ligand atom pairs. The general functional form is:

$$S_{ligand-target} = \sum_{i \in m, j \in n} \frac{p_{mat}(r_{i,j}^{P,L})'}{p_{ref}(r_{i,j}^{P,L})'} \quad (2.10)$$

where *i* and *j* are the atoms of the target and ligand, respectively; $r_{i,j}$ the interatomic distance between *i* and *j*; p_{mat} the distance distribution for a pair of atoms and p_{ref} the reference state for that pair of atoms. The distance distribution can be estimated from:

$$p_{mat}(r_{P,L}) \cong \frac{N(r_{P,L})}{\sum_{r_{\min} \le r \le r_{\max}} N(r_{P,L})} \quad (2.11)$$

where $N(r_{P,L})$ is the number of observations of atoms at a particular distance bin and r_{min} and r_{max} the minimum and maximum distances to be considered, respectively.

The reference state (P_{ref}) differs depending on the statistical potential. Here, we show the reference state of the scoring functions RankScore and PoseScore (Fan, Schneidman-Duhovny et al. 2011) implemented and used along the present work. In these functions, the reference state is calculated taking into account the frequency of the distances of all of the atoms:

$$p_{ref}(r_{P,L}) = \frac{\sum_{P,L} N(r_{P,L})}{\sum_{r_{\min} \le r \le r_{\max}} N(r_{P,L})} \quad (2.12)$$

where $\sum_{P,L} N(r_{P,L})$ is the number of all pairs of atom types in a particular distance bin.

To reach the final terms $p_{ref}(r_{i,j}^{P,L})'$ and $p_{mat}(r_{i,j}^{P,L})'$, the original terms are combined with the uniform distribution and the reference state respectively and the parameters are adjusted with a training set of complexes.

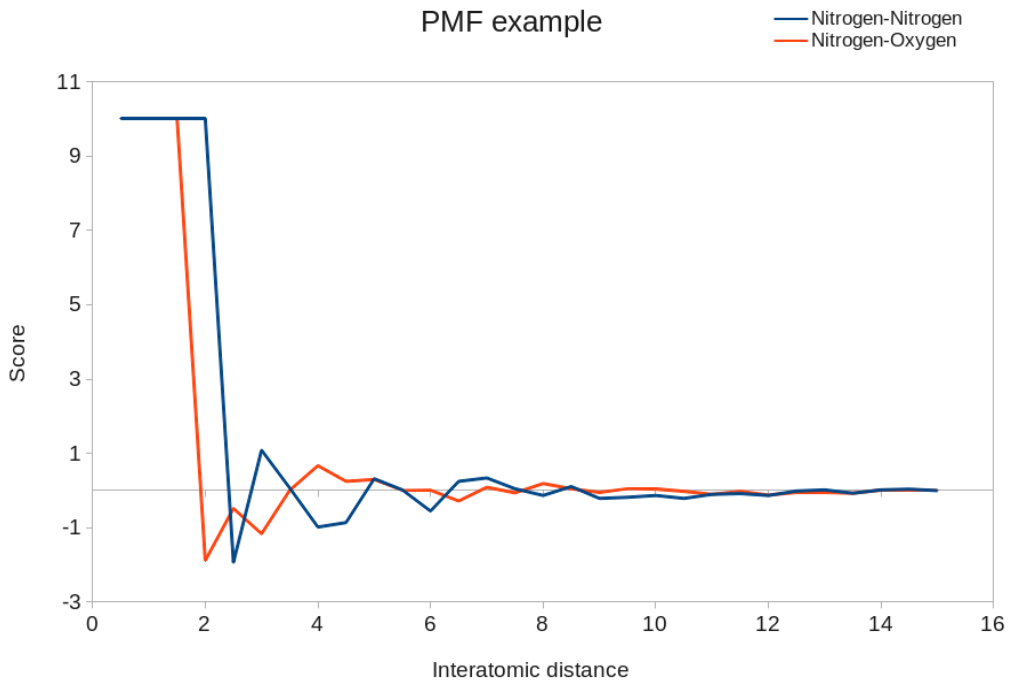


Figure 2.4. Potential of mean force for two different atomic pairs

2.6 Fingerprints and structural keys

Comparing ligands in an efficient manner is a difficult task. From the 80s there exist a large number of 2D and 3D comparison techniques that allows a fast screening of the increasingly vast chemical databases in just a few seconds (Eckert and Bajorath 2007). These versatile methods are grouped under the generic name of fingerprints.

Fingerprints are strings of 0s and 1s (bitstring) that encode a certain information about the ligand and can be processed really fast by computers due to their binary nature. The information encoded can be very different, from whole molecular fragments to the number of atoms or bonds and inter-atomic distances between functional groups.

The basic principle of fingerprint filtering or searching is that similar molecules (according to the selected criteria) should share activated bits in their fingerprints.

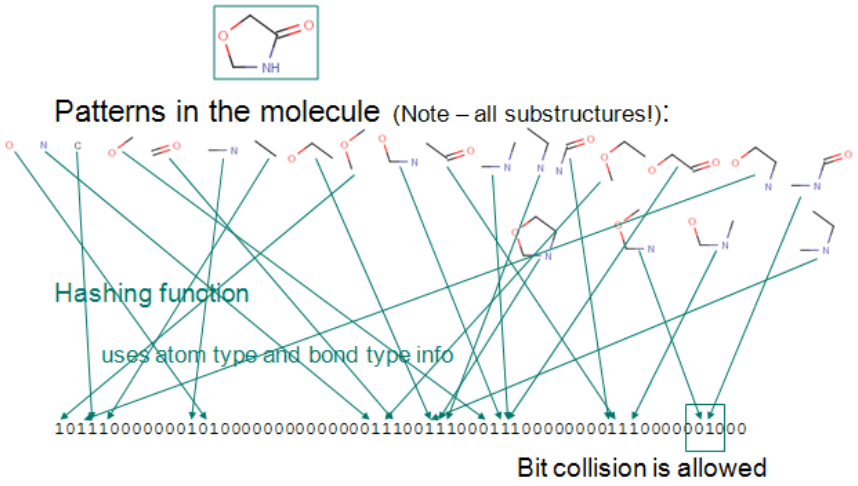


Figure 2.5. Fingerprint encoding that employs a hashing function for mapping fragments of 1 to 7 atoms in both directions. Adapted from ChemAxon's website(2013).

2.6.1 MACCS

(MDL) Molecular ACCess System is a structural key of 166-bits that indicates whether a predefined substructure or functional group is present or not. These fragments are designed to perform substructure and similarity searching in common molecular databases.

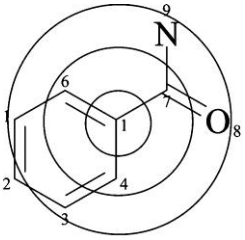
2.6.2 Group fingerprints

This fingerprint is defined by 306 different chemical groups (O'Boyle, Banck et al.) and allows filtering large databases by chemical moieties such as amides, carboxylic acids or benzyl derivatives.

2.6.3 Extended Connectivity Fingerprints (ECFP)

This family of topological (2D) fingerprints is based on the Morgan algorithm (Morgan 1965) which was initially designed to be able to recognize whether two molecules with different atom order are indeed the same molecule (molecular isomorphism).

The Morgan fingerprint assigns, in a first iteration, labels to the atoms based on their connectivity. In the following iterations it uses the previous labels to classify the atoms until each atom has a unique identifier. In ECFP this algorithm (Rogers and Hahn 2010) is modified to run only a certain number of iterations rather than to complete molecular disambiguation. Thus the intermediate labels in ECFP are the base of the fingerprint and not a temporal value as they were in the original implementation.



Considering atom 1 in benzoic acid amide

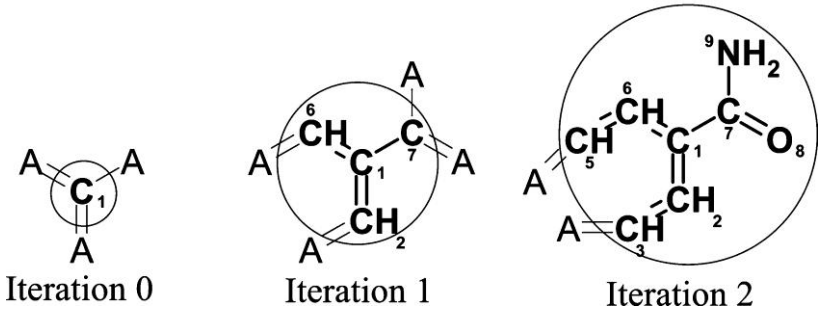


Figure 2.6. Numbering iteration process in ECFP. Adapted from Extended Connectivity Fingerprints(Rogers and Hahn 2010)

After each iteration, the values generated are used as input for a hash function that produces integer values uniformly in integer space (2^{32} in 32 bits). These are the base for the fingerprint. The space and the numbers generated are important to avoid collisions or the possibility that two different structures generate the same values.

2.7 Similarity metrics

2.7.1 Tanimoto coefficient

The Tanimoto coefficient (Rogers and Tanimoto 1960) is the most popular metric for similarity measurement using fingerprints and other techniques such as shape overlapping between two structures A and B. It can be defined as the coefficient between the intersection of A and B and the union of A and B:

$$Tc = \frac{|A \cap B|}{|A \cup B|} \tag{2.13}$$

It can also be expressed computationally in terms of the number of common and of all activated bits in two different bitstrings:

$$Tc = \frac{count\_bits\_on(a \,\&\, b)}{count\_bits\_on(a \mid b)} \tag{2.14}$$

2.7.2 Tversky index

This measure can be seen as a generalization of the Tanimoto coefficient (Senger 2009). It uses two parameters α and β to balance the weight of the features of template and candidate, respectively:

$$Ti = \frac{|A \cap B|}{|A \cap B| + \alpha|A - B| + \beta|B - A|} \tag{2.15}$$

where *α* is the coefficient for the template and *β* the coefficient for the candidate. Depending on the values for these two coefficients the Tversky index offers different alternatives:

- α = β = 1. This is equivalent to the Tanimoto coefficient since:

$$|A \cup B| = |A \cap B| + |A - B| + |B - A| \tag{2.16}$$

- α = β = 0.5. This is equivalent to the Dice's coefficient which will not be used in the present work.

- α = 1, β = 0. In this case, only the features of the template are taken into account and the filtering becomes a superstructure search where a maximum similarity of 1.0 means that the candidate has all the features included in the template and the minimum of 0.0 means that no features of the template are found in the candidate.

- α = 0, β = 1. This last case is particular useful for substructure search. As only the candidate features are important, the maximum values of 1.0 will be assigned to structures that have the template features embedded.

2.7.3 Manhattan distance

Some techniques to be introduced in following next sections produce a series of descriptors or float values to describe the molecular properties. Thus, to calculate the similarity in this '*property space*' it could be useful to use a metric derived from the Manhattan distance.

Let *p* and *q* be two descriptors vectors. To calculate the distance:

$$d(p,q) = \|p - q\| = \sum_{i}^{n} |p_i - q_i| \tag{2.17}$$

where *n* is the number of descriptors. The distance is not bound and Armstrong et al. (Armstrong, Morris et al. 2010) have proposed the normalized inverse metric:

$$dinv(p,q) = \frac{1}{1+\frac{1}{n}d(p,q)} \tag{2.18}$$

2.7.4 Root-mean-squared deviation (RMSD)

RMSD accounts for the average distance between the equivalent atoms in both molecules.

$$RMSD(A,B) = \sqrt{\frac{1}{n}\sum_{i}^{n}\left\|a_i - b_i\right\|^2} \tag{2.19}$$

where *A* and *B* are the coordinates of the equivalent atoms of the two molecules and *n* is the number of atoms.

It is usually employed in docking validation sets where the crystallographic solution is available and to assess the quality of both, algorithms and scoring functions. Acceptable values for RMSD depend on the application: for small molecular fragments it has a limit of 1.5 Å whereas for a normal drug-like molecule the limit is 2.0 Å. The main limitation of this measure is that small deviations of certain parts of the molecules (*e.g.* rings) can increase dramatically this value despite the fact that the key interactions determining the spatial configuration of the two entities are well reproduced. Several other measures have been proposed such as the Generally Applicable Replacement for RMSD (GARD) (Baber, Thompson et al. 2009) which addresses this fundamental issue. However they are still not very popular and for comparison purposes RMSD is still in use.

2.7.5 TM-score

Comparison of protein structures in the search for similar binding sites requires a robust metric that overcomes the limitations of the RMSD. The TM-Score algorithm (Zhang and Skolnick 2004) was designed to calculate the structural similarity between two protein models:

$$TMScore = Max\left(\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right) \tag{2.20}$$

where L_N is the length of the native protein, L_T the length of the aligned residues, d_i the distance between the *i*th pair of aligned residues, and d_0 a normalized scale with a value of 0.17. *Max* function denotes the maximum values after spatial superposition.

2.8 Pharmacophores

Pharmacophores are groups of electronic or steric features in a molecule that are required for binding a target or triggering its response. Most common features include hydrogen bond donors, hydrogen bond acceptors, positively charged groups, negatively charged groups, lipophilic groups and aromatic rings. The main advantage of these features is that they are very general in the sense that many different chemical groups can fulfill the requirements of the definitions, *e.g.* a negatively charged group may include a carboxylic acid as well as a tetrazole moiety. Pharmacophores can be defined by using the information available in ligands that are already known or derived from important spots in the binding site that a ligand should match.

2.8.1 3D pharmacophores

Traditional pharmacophores are usually defined as a group of chemical features in Cartesian space. The distances between specific features tend to be expressed in bins in order to improve the computational speed.

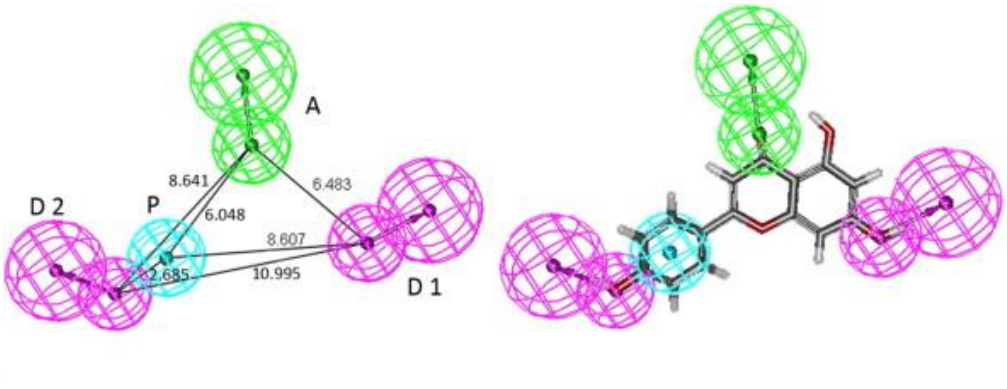


Figure 2.7. Pharmacophoric definitions and a compound that matches them (Liu, Zhou et al. 2012).

2.8.2 2D or topological pharmacophores

These pharmacophores encode the information of the relative position between features using bonds and atomic connectivity instead of an absolute spatial distance. The most commonly used ones are the Chemical Advanced Template Search (CATS) 2D pharmacophoric fingerprints. CATS were introduced by Schneider et al. (Schneider, Neidhart et al. 1999) and define five classical pharmacophoric types (positive, negative, acceptor, donor and lipophilic) and up to 10 topological distances (from 1 to 10 bonds between the two points) expressed in terms of up to 150 different flags that can be

activated or not in a given molecule. These flags can be translated from present/not present to 1 or 0 and grouped to form a bitstring. The resulting bitstrings can be compared or used as templates in virtual screening.

```
Positive_Positive1:
[+1,+2,+3,+4,NX3H2]~[+1,+2,+3,+4,NX3H2]
Positive_Donor1: [+1,+2,+3,+4,NX3H2]~[OX2H,NX3H1,NX3H2]
Positive_Aceptor1: [+1,+2,+3,+4,NX3H2]~[O,N!H]
Positive_Negative1: [+1,+2,+3,+4,NX3H2]~[-1,-2,-3,-
4,CX3O2,SX4O2H1,PX4O2H1]
Positive_Lipophilic1:
[+1,+2,+3,+4,NX3H2]~[Cl,Br,I,SX2C2,CX4H3$([#6]),$([C,c])
X4H2$([#6])$([#6]),$([C,c])X4H1$([#6])$([#6])$([#6]),$([
C,c])X4$([#6])$([#6])$([#6])$([#6])$([#6]),cc]
Negative_Negative1: [-1,-2,-3,-
4,CX3O2,SX4O2H1,PX4O2H1]~[-1,-2,-3,-
Negative_Donor1: [-1,-2,-3,-
4,CX3O2,SX4O2H1,PX4O2H1]~[OX2H,NX3H1,NX3H2]
Negative_Aceptor1: [-1,-2,-3,-
4,CX3O2,SX4O2H1,PX4O2H1]~[O,N!H]
Negative_Lipophilic1: [-1,-2,-3,-
4,CX3O2,SX4O2H1,PX4O2H1]~[Cl,Br,I,SX2C2,CX4H3$([#6]),$([
C,c])X4H2$([#6])$([#6]),$([C,c])X4H1$([#6])$([#6])$([#6]
),$([C,c])X4$([#6])$([#6])$([#6])$([#6])$([#6]),cc]
Donor_Donor1: [OX2H,NX3H1,NX3H2]~[OX2H,NX3H1,NX3H2]
Donor_Aceptor1: [OX2H,NX3H1,NX3H2]~[O,N!H]
Donor_Lipophilic1:
[OX2H,NX3H1,NX3H2]~[Cl,Br,I,SX2C2,CX4H3$([#6]),$([C,c])X
4H2$([#6])$([#6]),$([C,c])X4H1$([#6])$([#6])$([#6]),$([C
,c])X4$([#6])$([#6])$([#6])$([#6])$([#6]),cc]
Aceptor_Aceptor1: [O,N!H]~[O,N!H]
Aceptor_Lipophilic1:
[O,N!H]~[Cl,Br,I,SX2C2,CX4H3$([#6]),$([C,c])X4H2$([#6])$
([#6]),$([C,c])X4H1$([#6])$([#6])$([#6]),$([C,c])X4$([#6
])$([#6])$([#6])$([#6])$([#6]),cc]
Lipophilic_Lipophilic1:
[Cl,Br,I,SX2C2,CX4H3$([#6]),$([C,c])X4H2$([#6])$([#6]),$
([C,c])X4H1$([#6])$([#6])$([#6]),$([C,c])X4$([#6])$([#6]
)$([#6])$([#6])$([#6]),cc]~[Cl,Br,I,SX2C2,CX4H3$([#6]),$
([C,c])X4H2$([#6])$([#6]),$([C,c])X4H1$([#6])$([#6])$([#
6]),$([C,c])X4$([#6])$([#6])$([#6])$([#6])$([#6]),cc]
```

Figure 2.8. CATS SMARTS definitions. Only the first topological distance for each class is shown

2.9 Shape similarity methods

These methods rely on the key-lock paradigm which states that targets behave as a rigid lock with a predefined shape that the key (ligand) must meet to be able to bind or trigger the biological response. Although it was shown in the Introduction that this approach is far from reality and oversimplified, the truth is that these methods provide a useful starting point in the search for new compounds if there are other previously known ligands available. They do not require any structural information regarding the target and tend to perform scaffold-hopping better than do traditional methods based on fingerprints. In this work two methods of this kind were used: ElectroShape and Gaussian overlap.

2.9.1 ElectroShape

The ElectroShape method was developed originally by Armstrong et al. in 2010 (Armstrong, Morris et al. 2010) and it was designed as a four dimensional extension of a previous method called UltraShape Recognition (USR) (Ballester and Richards 2007) by adding the atomic partial charge as a new descriptor.

USR encodes the shape information using the first, the second and the third moments of the distributions of distances from five different points around the molecule (number of dimensions + 1). Using the distributions of distances ensures that these values are invariant to translation and rotation. These points (centroids) are defined according to the atom positions. The first centroid is the unweighted barycentre; the second one is the furthest atom from the first centroid; and the third centroid is the atom furthest from the second centroid. The forth centroid in USR is the atom closest to the first centroid but it is not used in ElectroShape, which replaces this centroid using the inner product of the vectors defined by the first three centroids and summing the first centroid to define two additional points.

To measure the similarity between molecules, the authors proposed the use of the inverse of the Manhattan distance normalized by the number of descriptors (15) which spans from 0 to 1. Thus totally dissimilar molecules score 0 and identical molecules score 1.

The advantages and limitations of this method are clear. On the one hand these distribution moments or descriptors can be calculated in advance and stored easily, thereby reducing the computational resources needed to perform virtual screening. Also it provides a radically different viewpoint of the molecular shape that does not require overlapping optimization. On the other hand, the method only provides a one-dimensional metric for assessing similarity and the results cannot be visualized.

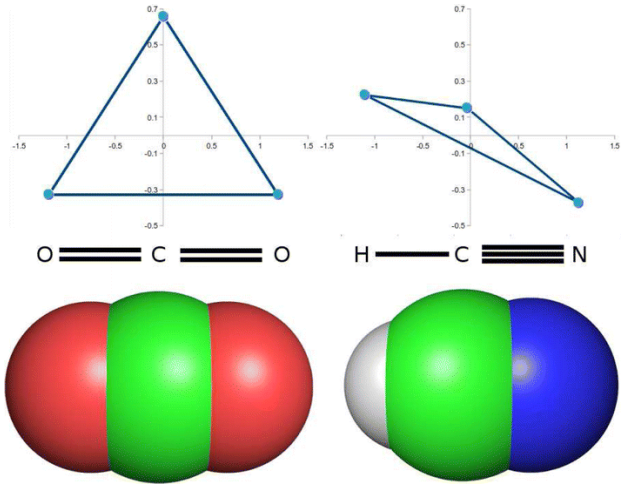


Figure 2.9. One-dimensional molecules carbon dioxide and hydrogen cyanide. These molecules are very similar in terms of steric shape, but look very different when the partial charges are added as an extra dimension. Adapted from Armstrong et al.(Armstrong, Morris et al. 2010).

2.9.2 Gaussian molecular overlap

This method uses the Gaussian shape model (Grant, Gallardo et al. 1996) where each atom is approximated by a spherical Gaussian function:

$$\rho_i^g(r_i) = p_i e^{(\alpha_i r_i^2)} \tag{2.21}$$

where r_i is the distance from the atomic center and α is defined by:

$$\alpha_i = \frac{\kappa_i}{\sigma_i} \tag{2.22}$$

where σ_i is atomic radius, with κ_i:

$$\kappa_i = \frac{\pi}{\sqrt[3]{\lambda^2}} \tag{2.23}$$

The Gaussian weigh (p_i) is selected to be:

$$p_i \lambda_i = \frac{4\pi}{3} \tag{2.24}$$

From parameter optimization, p_i = 2.70 and λ_i = 1.5514, and the values of the radius of each atom is adapted to its size. The calculation of the Gaussian overlap between two different atoms can be expressed as:

$$V_{ij} = p_i p_j \mathrm{K}_{ij} \left(\frac{\pi}{\alpha_i + \alpha_j} \right)^{3/2} \quad (2.25)$$

where *Kij* is:

$$\mathrm{K}_{ij} = e^{-\frac{\alpha_i \alpha_j r_{ij}^2}{\alpha_i + \alpha_j}} \quad (2.26)$$

2.9.3 Optimization algorithm

To calculate the shape similarity for two given molecules it is first necessary to optimize the overlap between them. This can be achieved with a variety of optimization algorithms. In the implementation developed in this work the downhill simplex method and the BFGS algorithm were used. These implementations are based on the analytical solution of the equations or are calculated in a pre-defined grid in which the template ligand is embedded. In our applications the downhill simplex method is activated by default, as it is more efficient in finding the optimum overlap from any arbitrary starting position and orientation.

2.9.4 Starting positions

Previous to the shape optimization step, both molecules are standardized in their spatial positions and orientations. First, their center of mass is translated to the origin (0,0,0). Then, the inertia tensor is calculated and diagonalized to obtain the rotation matrix to align the principal axes of the ligand with the reference frame. For the ligand, other three different starting orientations are calculated rotating 180 degrees each axis (Rush, Grant et al. 2005). The combination of these four different orientations and the downhill simplex algorithm make it highly probable to find the optimal overlap between the two molecules.

2.9.5 Overlap score

To measure the overlap, the Tanimoto index is employed and defined as:

$$T_c = \frac{O_{ab}}{O_{aa} + O_{bb} - O_{ab}} \quad (2.27)$$

where O_{ab} is the Gaussian overlap between molecules *a* and *b*, O_{aa} the self-overlap of the molecule *a* and O_{bb} the self-overlap of molecule *b*.

To improve chemical matching, an additional score is also implemented that only takes into account the overlap between chemically compatible atoms (*e.g.* hydrogen bond acceptors). The addition of this last Tanimoto score defines the total overlap measure that ranges from 0 (no similarity) to 2 (identity).

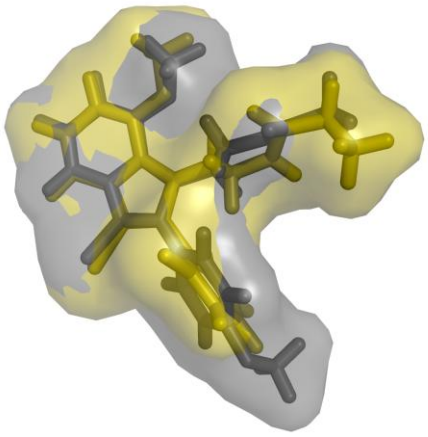


Figure 2.10. Shapes of Gaussian superimposed molecules.

2.10 Virtual Screening

There exist an increasing number of computational chemical databases that include most of the synthetically feasible compounds ever made in the world. However, and unfortunately, compound records rarely contain binding information or biological evaluation data.

Virtual screening is the massive application of the techniques already described (docking, fingerprints, shape similarity, etc.) to screen these databases with the aim of selecting compounds with a higher probability to bind to certain target or group of targets. These methods can exploit the information contained in structural targets or in known ligands.

2.10.1 Evaluating results

According to several reported analysis, the accuracy of virtual screening (VS) results appears to be extremely dependent on the target (Cross, Thompson et al. 2009; Armstrong, Morris et al. 2010; Cabrera, Klett et al. 2012), most probably due to the fact that the large number of approximations employed are not universally applicable. To evaluate any new VS methodology, a database should be screened for an unknown target and binding of the selected compounds to the intended target should be confirmed. Those

experimental validations are not always possible and for that reason there exist some *ad-hoc* datasets built from already known true binders to several targets and some decoys, which have been selected to mimic the physicochemical properties of these ligands. The most popular dataset is the Directory of Useful Decoys (DUD) (Irwin 2008) and this has been used throughout the work described in this thesis.

Several metrics can be applied to the problem of identifying active molecules from decoys (needles in a haystack):

2.10.2 Receiver operating characteristic (ROC) plot

ROC curves try to show the performance of a binary classifier. It plots 1 – specificity (or false positive rate) against sensitivity (or true positive rate). The area under the curve (AUC) is generally used as a metric for the global performance of the method. The main problem with the AUC is that the plot information is missed and the early recognition problem appears when the AUC is not able to distinguish between very good early recoveries and poor late recoveries since both would have the same overall AUC.

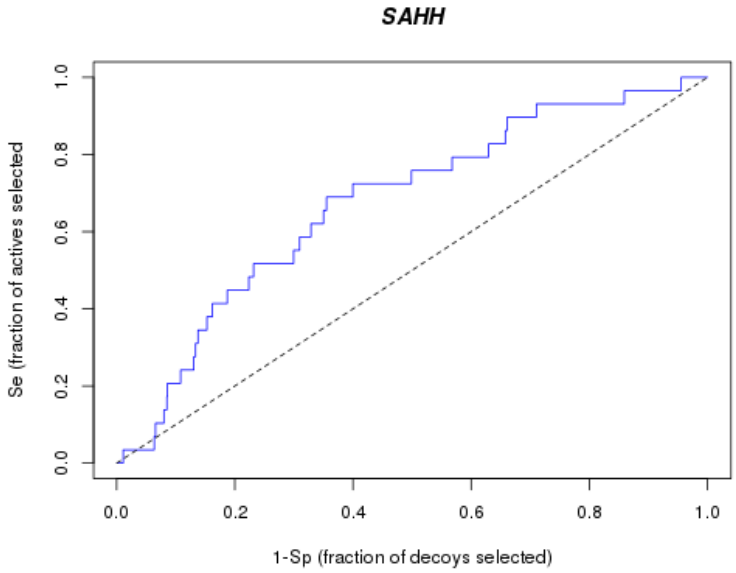


Figure 2.11. Example of ROC plot for a virtual screening experiment on adenosylhomocysteinase.

2.10.3 Enrichment factor

To solve the ROC AUCs problem the Enrichment factor at a given percentage (*N%*) of the database can be used. This metric measures the concentration of annotated ligand among the top N% of the virtual screening results compared with their concentration in the entire database.

$$EF_{N\%} = \frac{Ligands_{topN\%} / Compunds_{topN\%}}{Ligands_{total} / Compounds_{total}} \tag{2.28}$$

2.10.4 Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC)

The Enrichment factor is able to account for early recognition. However, it is useless for the rest of the database after the selected percentage. To solve both problems, namely enrichment factor and ROC AUC, Truchon *et al*.(Truchon and Bayly 2007) introduced the concept of BEDROC, which is similar to ROC AUC but it weighs top compounds more than the rest, including both results in a global value. The percentage of the database to overweight can be modified through a simple α parameter; the larger the value the more top compounds are taken into account.

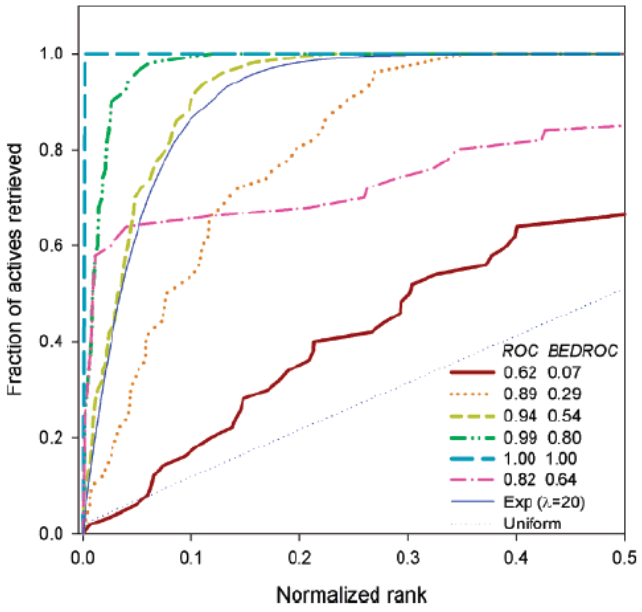


Figure 2.12. ROC curves with their corresponding AUCs and BEDROC AUCs. Adapted from Truchon *et al*.(Truchon and Bayly 2007)

2.11 Ligand efficiency indices (LEIs)

In the process of developing new molecules, there is a tendency to gain affinity for the target at the cost of increasing the size of the new entities. If good potency is necessary it is also required that the final compound has favourable physicochemical properties such as molecular weight, polarity and lipophilicity which ensure that the potency is retained *in vivo* when the compound is distributed through the patient's body.

The need to optimize several parameters at the same time together with the potency has led to the development of the ligand efficiency indices (Abad-Zapatero 2007) which normalizes the binding energy by the property to be optimized:

Name	Definition		Property
BEI	$\frac{\mathrm{pK_i\ or\ pK_d\ or\ pIC_{50}}}{\mathrm{Mw\,(in\ kDa)}}$	(2.29)	Weight
SEI	$\frac{\mathrm{pK_i\ or\ pK_d\ or\ pIC_{50}}}{\frac{^{1}\mathrm{PSA}}{100}}$	(2.30)	Polarity
NSEI	$\frac{\mathrm{pK_i\ or\ pK_d\ or\ pIC_{50}}}{^{2}\mathrm{NPOL}}$	(2.31)	Polarity
NBEI	$\frac{\mathrm{pK_i\ or\ pK_d\ or\ pIC_{50}}}{^{3}\mathrm{N}HEA}$	(2.32)	Weight
nBEI	$-\log_{10}\left(\frac{K_x}{^{3}NHEA}\right)$	(2.33)	Weight
mBEI	$-\log_{10}\left(\frac{K_x}{Mw}\right)$	(2.34)	Weight

Table X. [1]Polar Surface Area (calculated). [2]Number of polar atoms (N+O). [3]Number of non-hydrogen atoms.

These indices can be used as variables for 2D plots where the chemicobiological properties of the compounds are represented. The affinity is represented in the radial coordinate while the molecular weight and polarity are depicted in the angular coordinate.

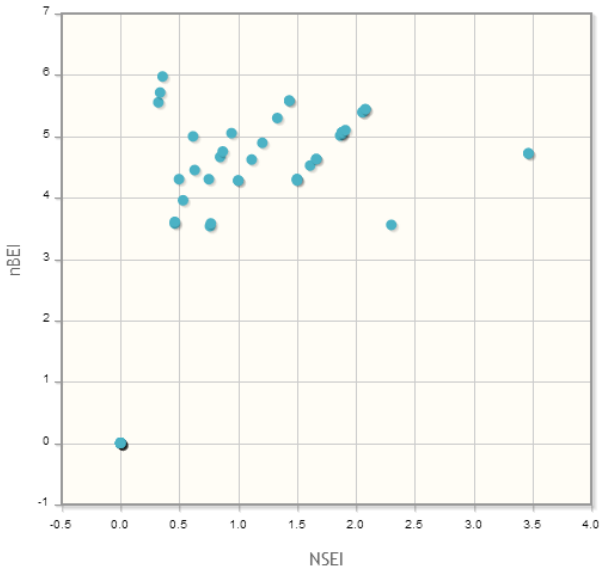


Figure 2.13. A 2D plot of NSEI and nBEI efficiency indices.

LEIs planes have been successfully used to track optimization pathways in retrospective analsysis (Blasi, Arsequell et al. 2011) where they are able to point out the most favorable modifications to the original compound towards a final drug candidate. In these optimization projects, molecules tend to be bloated with chemical features that not always contribute enough to the binding free energy to overcome the fact that they confer worse physicochemical properties and therefore a dismiss ability to act *in vivo*.

In addition, a framework based on LEIs could be useful to represent the contents of the databases that contain both chemical (compounds) and biological (activity) records such as BindingDB (Liu, Lin et al. 2007) or ChEMBL (Gaulton, Bellis et al. 2012).

3. OBJETIVOS

1. Desarrollo de una interfaz gráfica, métodos basados en ligandos y actualización de la plataforma de cribado actual VSDMIP.

2. Desarrollo de un nuevo protocolo de cribado virtual basado en fragmentos moleculares para acortar los tiempos de cálculo e incrementar la diversidad química de los resultados.

3. Actualización y mejora de la precisión y rapidez de cálculo de la herramienta CDOCK de cribado virtual basado en *docking* sobre las estructuras de las dianas macromoleculares.

4. Desarrollo de un servidor Web (AtlasCBS) para el análisis del espacio químico-biológico con aplicación en la optimización de compuestos (proceso hit-to-lead).

5. Desarrollar una herramienta para la predicción de efectos secundarios y análisis de redes en polifarmacología.

4. TRABAJOS DE INVESTIGACIÓN

Background and author's contribution

The was initially developed at the Bioinformatics Unit, Centro de Biología Molecular "Severo Ochoa", by Dr. Rubén Gil Redondo and colleagues, based on a previous development of a docking engine called CDOCK (Pérez and Ortiz 2001), which was written by Angel Ramirez Ortiz (and later improved by Dr. Gil), and several pieces of software developed to incorporate existing tools in a complete distributable platform, capable of being ported to several computational infrastructures such as the Bioinformatics Unit cluster, the Mare Nostrum supercomputer or the Berkeley Open Infrastructure for Network Computing (BOINC) distributed over standard PCs.

From the daily use of the platform and the analysis of the results produced, we identified several issues that could be improved and also multiple desirable features that were lacking in the 1.0 version, namely:

- **Speed**. The main docking engine, CDOCK, was based on a combination of exhaustive sampling and Monte Carlo Simulated annealing algorithms implemented in Fortran77. However, the optimization step following the initial sampling was based on a SIMPLEX downhill algorithm that is neither deterministic, which means that the solution to the problem could be arbitrarily missed, nor fast, which was a real issue for high-throughput VS.

- **Accuracy**. The scoring function implemented in CDOCK was entirely based on the non-bonding potential of the AMBER force field. This function, despite being able to discriminate extremely well native poses from non-native poses, was too restrictive and tended to discard near-native solutions generated by the previous sampling steps.

- **Lack of a ligand-based methodology**. The platform features were limited to filters based on protein structure (CDOCK, DOCK or Autodock) except for the module for ROCS. No filters based on physicochemical properties, such as the famous Rule-of-five, chemical moieties or ligand similarities were implemented.

- **Lack of ease of use**. The platform was based on command-line programs and XML configuration files that had be handled by the users themselves.

- **Flexibility**. All the protocols available considered ligand flexibility but neglected the inherent flexibility of the protein until the very end. Moreover, ligand flexibility was partially addressed by pre-generating a collection of plausible conformers according to energetic criteria but this proved to be not enough in some cases and to produce distorted geometries on a regular basis.

The objective of this part of the thesis was to overcome these limitations and to produce a new version of the platform (including a new docking engine) capable of improving the virtual screening results. The main author of these papers developed the new characteristics, tested the software and wrote the initial manuscripts.

Article I

VSDMIP 1.5: an automated structure-and ligand-based virtual screening platform with a PyMOL graphical user interface

J Comput Aided Mol Des (2011) 25:813–824
DOI 10.1007/s10822-011-9465-6

VSDMIP 1.5: an automated structure- and ligand-based virtual screening platform with a PyMOL graphical user interface

Álvaro Cortés Cabrera · Rubén Gil-Redondo · Almudena Perona · Federico Gago · Antonio Morreale

Received: 15 June 2011 / Accepted: 1 August 2011 / Published online: 9 August 2011

Abstract A graphical user interface (GUI) for our previously published virtual screening (VS) and data management platform VSDMIP (Gil-Redondo et al. J Comput Aided Mol Design, 23:171–184, 2009) that has been developed as a plugin for the popular molecular visualization program PyMOL is presented. In addition, a ligand-based VS module (LBVS) has been implemented that complements the already existing structure-based VS (SBVS) module and can be used in those cases where the receptor's 3D structure is not known or for pre-filtering purposes. This updated version of VSDMIP is placed in the context of similar available software and its LBVS and SBVS capabilities are tested here on a reduced set of the Directory of Useful Decoys database. Comparison of results from both approaches confirms the trend found in previous studies that LBVS outperforms SBVS. We also show that by combining LBVS and SBVS, and using a cluster of ~100 modern processors, it is possible to perform complete VS studies of several million molecules in less than a month. As the main processes in VSDMIP are 100% scalable, more powerful processors and larger clusters would notably decrease this time span. The plugin is distributed under an academic license upon request from the authors.

Á. C. Cabrera · F. Gago
Departamento de Farmacología, Universidad de Alcalá, 28871 Alcalá de Henares, Madrid, Spain

Á. C. Cabrera · R. Gil-Redondo · A. Perona · A. Morreale (✉)
Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Campus UAM, c/Nicolás Cabrera 1, 28049 Madrid, Spain
e-mail: amorreale@cbm.uam.es

Keywords Docking · Virtual screening · Drug design · Graphical user interface

Introduction

Many changes in the drug discovery paradigm have emerged in recent years due to major advances in the field of Computer-Aided Drug Design (CADD), which has benefited enormously from astounding improvements in the power of computers and new algorithms. As a result, attempts continue to be made to turn the drug discovery process into a more rational approach that can help design therapeutically relevant New Molecular Entities (NME) with a minimum of synthetic effort. Another important factor that has to be taken into the complex drug making equation is the vast amount of experimental information emanated from genome sequence and structural biology projects, as well as biochemical and biophysical studies, that is stored in publicly accessible databases. Precisely because of this complexity, CADD appears to be placed, at least conceptually, in an excellent position to help reduce the cost and time that it takes to launch a NME onto the market (a thousand million dollars and 15 years on average, respectively [1]). However, despite some signs of promise, the real results still fall far below the expectations [2].

CADD approaches include structure-based (SB) and ligand-based (LB) virtual screening (VS) of chemical (and fragment) libraries, and both methods are widely used in industry and academia. SBVS uses docking tools with the aim of identifying possible hits that can then be subjected to lead optimization. To this end, they are routinely tested for their capacity to (a) reproduce the experimental structures of a series of ligands bound to their receptor targets,

as found in high-resolution X-ray crystal structures, and (b) discriminate between true binders and fake ligands ("decoys") on the basis of a scoring function that, although far from accurately representing the free energy of binding that can be measured experimentally [3], is used to predict the strength of the receptor-ligand association. If the docking engine and the scoring function perform reasonably well in this respect, one can expect some success in the identification and ranking of putative hits in a VS experiment. As an alternative, and particularly in cases where the receptor's 3D structure is not available, it is also possible to use the geometry of one or more ligands that display affinity for this receptor as a query to try and fish out similar molecules from commercially available catalogues or databases. The selected compounds can then be tested experimentally for confirmation of affinity/activity.

Besides the core SBVS and LBVS algorithms, an integrated platform for VS studies needs some other pieces of software, the most important being those required for setting up receptors and ligands at the beginning of the procedure and for processing the results at the end. In addition, large databases are usually filtered according to some custom-made rules. The integration of all of these tools into a common, flexible, and user-friendly platform requires a great deal of effort because a series of *connectors* have to be developed to handle the existing variety of file formats. Besides, an adequate database engine needs to be used to store and process efficiently the massive amounts of data that are generated, in common with trends observed in other computational biology areas [4].

The growing interest in this type of computational platforms that put together all the essential pieces to enable the effortless execution of complex VS protocols has resulted in a number of applications. Some solutions are commercial, like the Schrödinger [5] and Sybyl [6] suites and Pipeline Pilot from Accelrys [7, 8]), but open-source plugins for the popular molecular graphics program PyMOL [9] have also been designed and released, e.g. the intuitive and user-friendly interfaces to widely used software such as AutoDock/Vina [10] or AMBER [11]. Furthermore, other implementations are accessible through a web server (DOCKBLAST [12]), distributed over a grid [13], or endowed with database capabilities [14]. In our lab, VSDMIP [15] was developed to provide the scientific community with a flexible, fully automated computational platform to perform VS experiments and manage every piece of data in an integrated fashion. Significant advantages of this platform are its underlying database, which stores ligand information and every result arising from the different steps of a given VS protocol, and its modular and pluggable architecture, which allows customization of each step of the procedure. However, the original VSDMIP only allowed SBVS to be done and worked through a command-line interface.

In this paper we describe the improvements that have been incorporated into the updated version (VSDMIP 1.5) to overcome these shortcomings: (1) an LBVS module has been built that can be used not only in cases where the receptor's 3D structure is unavailable but also as a complement to SBVS applications, and (2) a graphical user interface (GUI), written in Python programming language, that allows its facile use as a plugin to the popular molecular visualization program PyMOL. VSDMIP 1.5 is compared to similar existing software and its LBVS performance on a subset of the Directory of Useful Decoys (DUD) database [16] is reported. As with the original VSDMIP, we are committed to making this updated and more powerful platform available free of charge to academic and non-profit organizations so that the scientific community, and eventually society at large, can benefit from it.

VSDMIP extensions

LBVS: the newly added functionality

Molecular fingerprints

Molecular fingerprints are strings of bits denoting the presence (1) or absence (0) of certain types of molecular information, typically chemical groups or relevant interaction points. They can be 2D or 3D depending on the structural information encoded. VSDMIP allows the user to work with both 2D (MACCS [17], CATS [18], and chemical groups for filtering) and 3D fingerprints (triplets of interaction points).

2D fingerprints The Molecular Design Ltd. (MDL) Molecular ACCess System (MACCS) structural fingerprint is a 166-bit string that indicates whether a predefined substructure or functional group is present or not.

The Chemical Advanced Template Search (CATS) fingerprint is composed of a bit for each possible combination of hydrogen bond donor, hydrogen bond acceptor, positively charged group, negatively charged group, and lipophilic PharmacoPhoric Points (PPP), including aromatic rings, separated by distances between 1 and 10 bonds and totalling a length of 150 bits.

Finally, a group fingerprint is a 306-bit string that denotes the presence or absence of different chemical fragments and functional groups (see Open Babel documentation for details [19]). The use of this fingerprint is indicated as a post-filter after using MACCS or CATS to ensure that the selected compounds do possess the required functional groups.

In the three cases, the fingerprints can be calculated either from a given database or from a file containing a set of Simplified Molecular Input Line Entry Specification (SMILES) strings [20] using Open Babel [19], which is

also integrated within VSDMIP. In the database the fingerprints are stored in the FINGERPRINT table (Fig. 4). Also, a previously saved file containing molecule IDs and fingerprints can be loaded and stored within the database.

3D fingerprints 3D fingerprints can be defined using the 3D molecular structure and six types of PPP with the following interaction properties: hydrogen bond acceptor, hydrogen bond donor, positively charged group, negatively charged group, aromatic ring, and lipophilic point. These PPP, associated in triplets, can be automatically calculated for all the molecules in a given database. The generator of triplets (GTP) code recognizes all possible PPP for each conformer in the database and builds the triplets in hexadecimal strings representing the type and the relative disposition of the PPP. This information is stored in the PHARMACOPHORES table (Fig. 4). Additionally, the user can create customized 3D fingerprints by choosing the type of points on the graphical interface and moving them to a desired position. This fingerprint can then be used as a pattern to search for molecules in databases that fulfil these conditions. Finally, the molecules obtained from a search can be incorporated directly into the main workflow of SBVS.

Fingerprint comparisons

VSDMIP incorporates three coefficients for fingerprint comparison (Tanimoto, Tversky, and rule-based) as well as two mechanisms to combine queries (hybrid fingerprints and scoring fusion).

Tanimoto coefficient Given two objects, A and B, represented as two strings of bits, the Tanimoto coefficient, *Tc*, is defined as the ratio (Eq. 1):

$$Tc = c/(a + b + c) \quad (1)$$

where a is the count of *on* bits in object A but not in object B, b is the count of *on* bits in object B but not in object A, and c is the count of *on* bits in both objects A and B

It can be viewed as the ratio of *on* bits shared by both string representations. The values range between 0 (no similarity at all) and 1 (identical fingerprints).

Tversky coefficient It introduces the concept of a *prototype* to which the objects or *variants* are compared to and is defined as the following ratio:

$$T_v = c/(\alpha * a + \beta * b + c) \quad (2)$$

where a, b, and c have the same meaning as before, and α and β are weighting factors for the *prototype* and the *variant* so as to customize the relative importance of one *versus* the other. The T_v coefficient is also bound between 0 and 1.

Rule-based coefficient This is a ratio calculated as the result of the AND operation between the query string and the comparing counterparts, and therefore takes into account only those bits that are activated in the query string.

Mechanisms to combine queries: hybrid fingerprints versus scoring fusion

When more than one query is at hand (i. e., several known actives, to select some specific characteristics, or inactives, to rule out other non-desired properties), two options are available: (a) to combine the queries themselves, or (b) to combine their individual results. For the former, VSDMIP implements the centroids module, which analyses a set of query compounds and generates a fingerprint that concentrates all the information present in the whole set. A bit is activated if it is already present in a certain percentage of the compounds (the cut off is 0.5 by default but can be adjusted manually). Centroids can then be used like a regular fingerprint to query the database. For the latter, once multiple searches have been performed, common fusion scoring schemes are employed over the individual scoring values obtained: maximum, minimum, product, average, and the sum of the scores. These schemes have been implemented via a user-defined function (UDF), and it should be possible to include new operations easily. Finally, combined queries of active and inactive compounds could be useful for detecting ambiguous molecules: those giving good results when looking for actives and also good results when looking for inactives.

VSDMIP graphical user interface

Considering the advantages, for non-expert users, of incorporating a simple and easy-to-use interface to interact with the VSDMIP platform, we decided to create a plugin for the popular and versatile PyMOL graphics program as a tool to control the complete VS workflow (either LBVS or SBVS, or any combination thereof).

Once the plugin is invoked from the PyMOL plugin interface, it displays a simple window with a menu bar containing six submenus, the description of the program and the version number (Fig. 1). The menu bar includes: SBVS, LBVS, local operations (non-dependent on the database facilities, Local), database information (Info), configuration (Config) and information about the program, the authors, and contact details (About).

Under the SBVS submenu available options are:

(a) Insert new molecules, to incorporate new SMILES strings into the database,
(b) Run Virtual Screening, to create a new VS job and submit it to a cluster or to a multiprocessor machine,

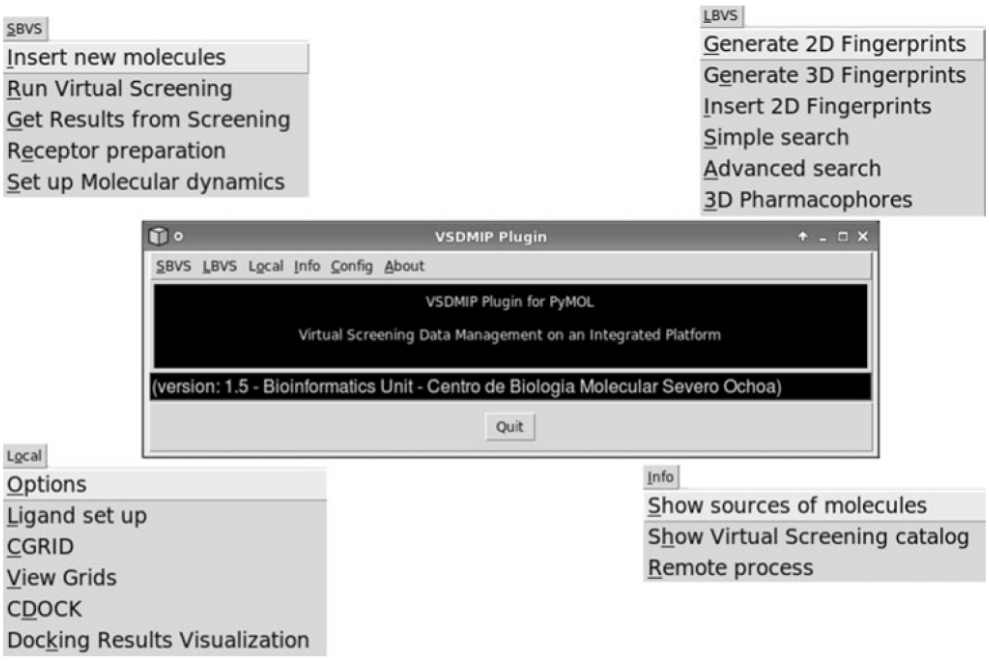


Fig. 1 VSDMIP's main window (*centre*) as it is launched from PyMOL, and windows that originate from the tabs that give access to the SBVS, LBVS, Local and Info tasks

(c) Get Results from Screening, to extract the results from a previous screening (structures, energy values, and general information about the interactions),
(d) Receptor preparation, to simplify the protein preparation workflow through a local script, and
(e) Set up Molecular Dynamics, to generate the topology [top] and coordinates files [crd] that are necessary to run a molecular dynamics simulation using the AMBER suite (http://ambermd.org/) [21].

The procedure for the latter operation, which makes use of AMBER's antechamber module, involves the mandatory creation of the ligand-related parameter and connectivity files [frcmod] and [prepin], the immersion of the complex in a user-adjustable cube of Transferable Intermolecular Potentials 3 Point (TIP3P) water molecules [22], and the addition of any necessary counterions to achieve electroneutrality.

The LBVS submenu gives access to the novel options for performing complete VS experiments using fingerprints of different types:

(a) Generate 2D Fingerprints, calculation of 2D fingerprints;
(b) Generate 3D Fingerprints, to generate triplets of PPP using the automated tool runGTP;
(c) Insert 2D Fingerprints, to insert the generated fingerprints into the database;
(d) Simple Search, to perform a single search using simple parameters;
(e) Advanced Search, to use special techniques for performing searches like scoring fusion, substructure search or chemical group filtering; and
(f) 3D Pharmacophores, to generate, within the PyMOL GUI, newly defined PPP to be used in new searches.

The Local submenu allows the user to carry out a complete docking process, from the conformational analysis of the ligand to the final visualization of the results. The available options are:

(a) Options, to configure the paths of the different programs;
(b) Ligand Set Up, to prepare the ligands for ALFA calculations (conformational analysis) and atomic charge assignment;
(c) CGRID, to calculate the energy grids for docking with CDOCK;
(d) View grids, to visualize the grids;
(e) Grid processing, to post-process the interaction energy grids by Boltzmann averaging (as a way to include receptor flexibility for docking [23]) or by calculating a grid as a difference of two other previously calculated grids (for example, on two related targets, as a way to explore selectivity [24]);
(f) CDOCK, to set up the docking process; and
(g) Docking Results Visualization, to analyse docking results (different energy terms and the type of interactions).

The Info submenu contains Show source of molecules and Show VS catalogue, two windows in which the user can look up information regarding the putative ligands and the VS protocol, respectively, as they are stored within the database (to be used later on as part of MySQL queries); and Remote process, to monitor the processes for which

instructions have been issued (the exact command is set up in the Config submenu).

The Config submenu contains the Configure queues tab that gives access to a window for defining the paths for the ssh and scp protocols, remote system and commands, and the MySQL settings.

Finally, the About submenu displays the name and affiliation of the authors, a contact address for further information, and the copyright statement.

Figure 2 illustrates the implementation of the VSDMIP GUI in PyMOL and how results from a docking run with CDOCK [25] can be visualized.

A case study

The test set: directory of useful decoys database

Eighteen targets (ACE, MR, HIVPR, P38 MAP, HMGA, PNP, COMT, Thr, TK, fXa, AChE, HSP90, COX1, COX2, AMPC, ALR2, GPB, and ER_a) were selected from the DUD database [16] ensuring enough diversity of types. Their bound ligands were downloaded directly from the original site and processed according to our established protocol. In short: a) conversion of all compounds into their isomeric SMILES [26] strings (to meet with the defaults in the VSDMIP protocol as described in the original work) and insertion into the database, which implies their transformation from 2D to 3D with CORINA [27], assignment of atom point charges using the Austin Model 1 electrostatic potential (AM1/ESP) fitting method, as implemented in MOPAC 7 [28], addition of AMBER [21] atomic radii, and conformational analysis with ALFA [29]; and b) calculation of CATS, MACCS and group fingerprints with a modified version of the Open Babel program (decimal output and a parser for CATS were added), and insertion in the extended database. Importantly, this processing means that the original geometry of the bound ligand is lost and that each molecule will be present in the database as a collection of ready-to-dock conformers. For the protein targets PDB2PQR [30] was used to assign AMBER force field atomic radii and charges, while the protonation states of titratable residues were decided on the basis of pK_a calculation carried out with the PROPKA software [31]. The binding site to be explored was delimited in each case by the location of the bound ligand in the X-ray crystal structure using CGRID [25].

Virtual screening

In LBVS, we have used the searching functions described above to query the database multiple times (as many as the number of actives) and retrieve similar compounds taking the Tanimoto coefficient as the score. The global performance for a given target was evaluated as the mean (over all its actives) of the area under the curve (AUC) values from receiver operating characteristic (ROC) plots, as well as the standard deviation of ligand atoms from their experimentally determined location. Centroid calculations

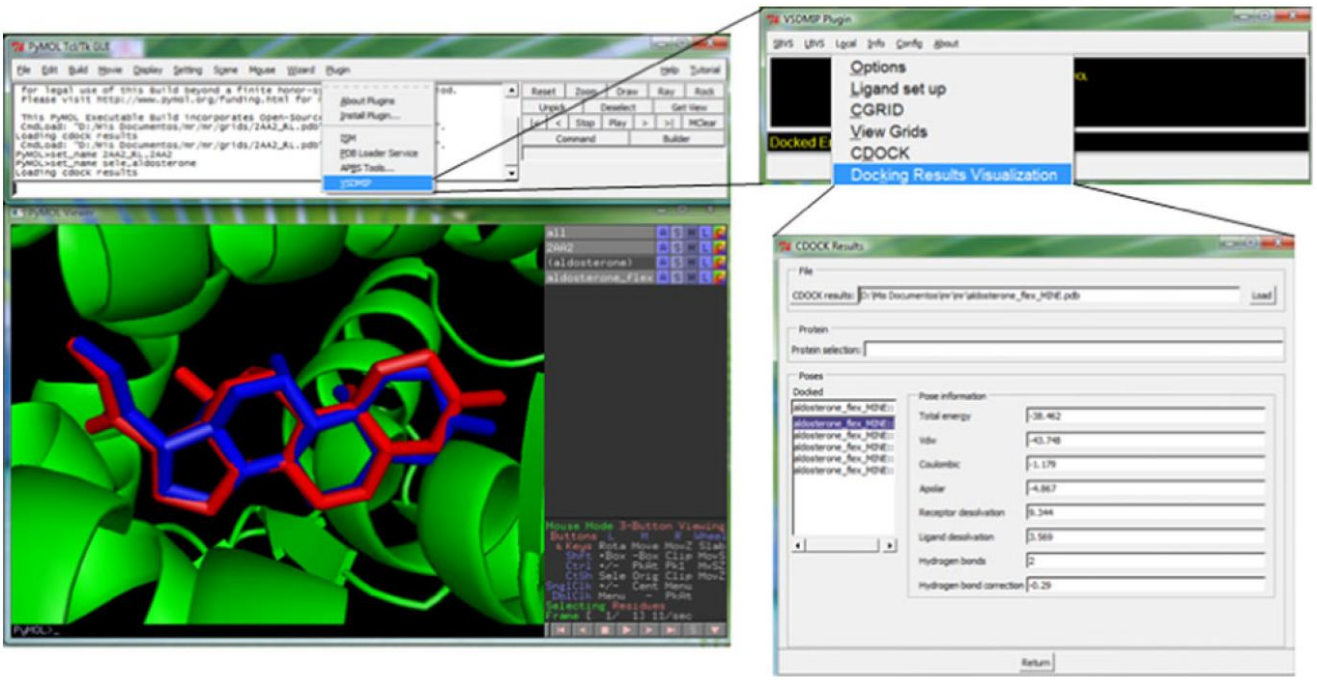


Fig. 2 Screenshot showing the interface between PyMOL (*left*) and VSDMIP (*right*). A CDOCK window displays the docking results for aldosterone (*blue sticks*) in the ligand-binding site of the mineralocorticoid receptor (*green*) compared to the X-ray crystal structure (*red sticks*). The *lower-right* window lists the binding energy values associated with each pose

were also performed for each target. For fusion calculations, 3 randomly chosen actives were selected in each round, and 3 rounds were performed for each target. Mean values and standard deviations are also reported.

SBVS was performed by the combined use of CGRID and CDOCK [25] as explained in the original VSDMIP publication [15]. Briefly:

(a) for each protein structure, the initial binding site was defined as the space delimited by the axis-parallel box containing the co-crystallized ligand, augmented by 5 Å in each axis direction;
(b) CGRID calculation of protein interaction fields (a 12–6 Lennard-Jones term and an electrostatic term modeled with a sigmoidal dielectric screening function) covering the binding site (0.5 Å spacing in all directions) using common atom probes (C, N, O, S, P, H, F, Cl, Br, and I);
(c) exhaustive exploration by CDOCK of the location and orientation of each molecule within the binding site by positioning their centers of mass on grid points and performing discrete rotations of 27° on each axis;
(d) energy evaluation of each pose by the molecular mechanics force-field scoring function implemented in CDOCK that can additionally include ligand and receptor desolvation energy terms as well as counting of hydrogen bonding interactions; and
(e) selection of the best-scoring pose for each molecule as the docking solution.

Finally, LBVS was employed prior to SBVS to illustrate the connection between both modules and also to test its ability as a filter to reduce docking times using ACE and MR as the targets. To this end, MACCS fingerprints were calculated for ACE and MR active ligands, and the centroids method was used to combine all the information for the actives into single queries that were employed to search the entire DUD database of decoys and retrieve only those with a $Tc > 0.6$. The selected molecules were then docked into their respective targets as explained above.

The results from LBVS and SBVS were evaluated using ROC plots [32], which represent sensitivity (y-axis, true positives rate, Eq. 3) versus specificity (x-axis, false positive rate, Eq. 4). AUCs were calculated for each ROC plot.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3)$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}. \quad (4)$$

Case study results: performance and timing

Using the eighteen targets selected from DUD we compared the performance of LBVS (through topology-based [MACCS] and mixed topological-physico-chemical-based [CATS] fingerprints) and SBVS (using docking and pharmacophore methods) as measured by the AUC values (see Table 1). For MACCS the centroids and fusion approaches were used to handle more than one initial query, whereas for CATS only the centroid approach was employed for testing purposes. In general, and considering the 18-target DUD reduced set, MACCS afforded the best results in terms of averaged AUC values (0.74 ± 0.17), followed by CATS (0.64 ± 0.11) and docking (0.63 ± 0.13) when only one query was used to search the database (only applicable to MACCS and CATS). No significant changes were observed in MACCS (0.72 ± 0.14) or in CATS (0.65 ± 0.10) when different queries were considered via the centroids method. A slight improvement was found for MACCS and the fusion method (0.79 ± 0.19), although only 6 cases were studied. Finally, the PPP method also performed reasonably well, but since it was applied to only two cases no definite conclusions can be drawn. A more detailed analysis can be done by splitting the AUC into three ranges: a lower-bound range where the AUCs are $\leq$0.5 (worse than random), 0.5–0.7 (above random but with room for improvement), and $\geq$0.7 (clearly better than random). On the one side, clearly better than random, MACCS afforded 60–70% of the targets with AUCs above 0.7, while these percentages were 30–40% for CATS and 20% for CDOCK. On the other side (worse than random) all the methods performed similarly (6%) although some variability was obtained for MACS depending on the number of starting queries (data not shown). In the middle range ($0.7 \leq \text{AUC} \leq 0.5$), CDOCK yielded the highest value ($\sim$70%), then CATS (40–60%), and finally MACCS (20–40%).

As will be discussed below, *analogue bias* can cause artificial enrichment because, if the query molecules are topologically similar to the actives, these will be retrieved more easily. To test this, we calculated similarity matrices using *Tc* among the actives for each target, and some examples are shown in Fig. 3 (HMGR, TK, COX1, and ALR2).

The use of LBVS (in this case MACCS) as an initial filter to a more computational demanding task such as docking resulted in an important reduction in computer time while maintaining the overall performance. For ACE, the number of docking experiments that had to be done after filtering out with LBVS was reduced from 1791 to just 422, and this meant a reduction of 77% in computation time. The AUC was 0.66, slightly above the value for docking alone. Similar results were also found for the MR target. In this case, a reduction of 74% in computing time was achieved by reducing the number of ligands for docking from 637 down to 163, while the AUC was improved in 0.1 units.

Table 1 AUC values for the two VS techniques (LBVS and SBVS) and the different methods according to each target studied

Target	LBVS					SBVS	
	MACCS			CATS			
	Single FP	Centroids	Fusion (average)	Single FP	Centroids	CDOCK	PPP
ACE	0.73 (0.09)	0.78	0.79 (0.48)	0.61 (0.13)	0.56	0.63	
MR	0.72 (0.22)	0.82	0.86 (0.03)	0.51 (0.09)	0.56	0.75	0.63
HIVPR	0.58 (0.13)	0.62		0.63 (0.12)	0.46	0.25	
P38 MAP	0.72 (0.16)	0.87	0.81 (0.10)	0.53 (0.20)	0.73	0.50	
HMGR	0.85 (0.22)	0.63	0.51 (0.46)	0.79 (0.09)	0.77	0.26	
PNP	0.85 (0.06)	0.71	0.91 (0.04)	0.62 (0.09)	0.63	0.60	
COMT	0.86 (0.13)	0.96		0.75 (0.21)	0.86	0.33	0.79
Thr	0.67 (0.15)	0.76		0.83 (0.11)	0.62	0.33	
TK	0.88 (0.05)	0.70		0.81 (0.11)	0.61	0.60	
fXa	0.77 (0.18)	0.61	0.88 (0.01)	0.83 (0.17)	0.60	0.55	
AChE	0.63 (0.10)	0.59		0.51 (0.07)	0.55	0.73	
HSP90	0.77 (0.62)	0.39		0.63 (0.06)	0.54	0.65	
COX1	0.48 (0.16)	0.63		0.50 (0.04)	0.67	0.53	
AMPC	0.90 (0.11)	0.70		0.72 (0.08)	0.66	0.47	
ALR2	0.57 (0.08)	0.62		0.47 (0.11)	0.61	0.33	
COX2	0.85 (0.41)	0.80		0.60 (0.13)	0.81	0.67	
GPB	0.81 (0.14)	0.89		0.72 (0.12)	0.72	0.83	
$ER_{agonists}$	0.75 (0.12)	0.92		0.48 (0.05)	0.70	0.63	

Numbers in parenthesis are the standard deviations

ACE angiotensin-converting enzyme, *MR* mineralocorticoid receptor, *HIVPR* HIV protease, *P38 MAP* P38 mitogen activated protein, *HMGR* hydroxymethylglutaryl-CoA reductase, *PNP* purine nucleoside phosphorylase, *COMT* catechol *O*-methyltransferase, *Thr* thrombin, *TK* thymidine kinase, *fXa* coagulation factor Xa, *AChE* acetylcholinesterase, *HSP90* human heat shock protein 90, *COX-1* cyclooxygenase-1, *AMPC* AmpC β-lactamase, *ALR2* aldose reductase, *COX-2* cyclooxygenase-2, *GPB* glycogen phosphorylase γ, $ER_{agonists}$ estrogen receptor (agonist-bound conformation)

To assess the applicability of VSDMIP to large-scale VS projects, based on ligands and/or receptors, we measured the overall performance of VSDMIP when undertaking the major tasks that are common to all the protocols such as inserting the molecules into the database, generating fingerprints, searching within the database, and docking (Table 2).

According to the data compiled in Table 2, an average of 370 and 275 molecules can be inserted and docked, respectively, per day and CPU using VSDMIP. The number of inserted molecules showed a very high correlation with the number of conformers per molecule ($r^2 = 0.98$, after exclusion of COMT due to the fact that the ligands for this target are few and very small, and therefore uncharacteristic of the most typical real world scenario). The correlation was more modest ($r^2 = 0.79$) when the number of heavy atoms, the number of conformations per molecule, and the number of valid grid points were simultaneously considered. As expected, LBVS proved to be several orders of magnitude faster than SBVS. In fact, VSDMIP is able to generate around 10^7 2D- (molecules) and 10^6 3D-fingerprints (conformers) per day and CPU. On the other hand, 10^9 and 10^6 comparisons can be performed using 2D- and 3D-fingerprints, respectively, per day and CPU.

Technical issues

The PyMOL plugin

The VSDMIP plugin implements a visual interface to manage the most common tasks in VS. The menu bar has three main categories of actions: SBVS, LBVS and Local.

SBVS holds a visual interface for the original workflow implemented in the first release of VSDMIP. It has been extended to use results originating from its LBVS counterpart and to perform docking on the receptor with the selected molecules.

LBVS encompasses database-related operations (such as filters and search tools) to perform similarity calculations on 2D/3D fingerprints (codifications of several molecular features) and does not require the 3D structure of a receptor.

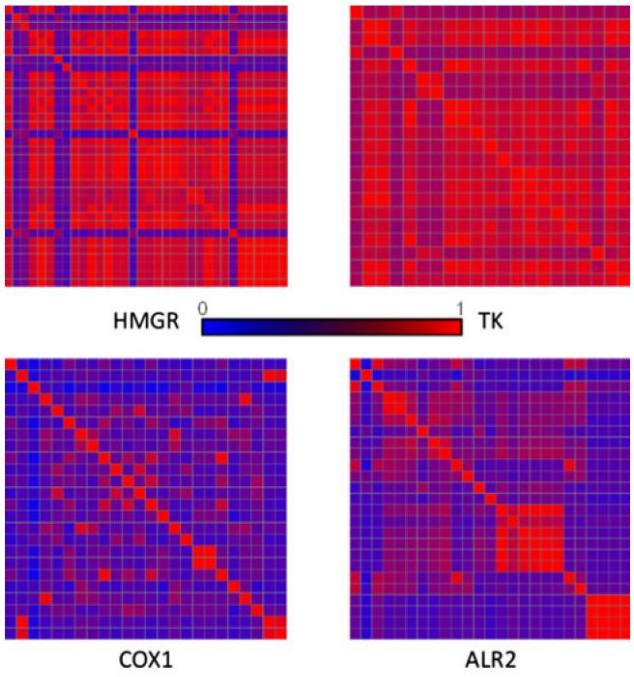


Fig. 3 Similarity matrices using Tanimoto coefficients for true binders to four different targets. Note that the *analogue bias* effect is clearly shown on HMGR and TK, while it is completely absent in COX1 and ALR2

Local allows the user to execute a complete docking job on a local machine without extending the capabilities of the basic programs (database-related functionalities). The SBVS and LBVS modules are designed to work within a cluster of processors or on a multiprocessor machine. The main node on the cluster communicates with the other calculation nodes through the secure shell (SSH) protocol in a way that is completely transparent to the user.

LBVS techniques have been implemented that extend currently available tools (Open Babel [33], MySQL, and PyMOL) and new programs and interfaces have been designed, e.g. the GTP code to enable the recognition of all possible PPP (see above at the "3D fingerprints" section).

The MySQL Application Programming Interface (API) was used to build User Defined Functions (UDF) in order to hold the methods for comparing and scoring fingerprints, including scoring fusion. The UDF are loaded directly into the main memory whenever the database requires them so that the filtering and searching processes are speeded up.

Extensions to VSDB

Three new tables have been added to the Virtual Screening DataBase (VSDB) contained in the original VSDMIP: FINGERPRINT, PHARMACOPHORE, and FINGER_TYPE (see Fig. 4). The first and second tables contain 2D and 3D fingerprints, respectively, whereas the third one stores the information and description of the available fingerprint types. The original role played by VSDB in the previous version of VSDMIP (storing molecules and results) has been expanded with filtering and searching tools within the MySQL engine. To this end the MySQL language was complemented with new functions that allow advanced molecular screenings based on similarity calculations to be performed. Backward compatibility with the original VSDMIP is maintained, as the plugin developed here is an optional upgrade.

Operating system and software/hardware requirements

The client version of the platform is compatible with the Linux and Windows operating systems. However, the

Table 2 Some ligand- and binding site-related properties of the complexes studied here and overall VSDMIP performance in the main operations

	Target[a]	⟨NHA⟩[b]	⟨Conf⟩[c]	BSGP[d]	Insertion	Docking			2D	3D
SBVS	ACE	23	118	18225	192	336	LBVS	FP Generation[e]	9×10^6	1×10^6
	COMT	16	35	7956	6600	288				
	PDE5	30	136	7182	192	168				
	AChE	26	133	7040	200	342		Search	2.4×10^9	0.5×10^6
	PARP	20	16	4332	1080	288				
	Thr	32	136	2688	192	221				

For structure-based VS (SBVS), the Insertion and Docking columns display the number of molecules that are processed per day and CPU (either a PIV 32-bit 3.2 GHz or a Xeon 32-bit 3.06 GHz processor). For ligand-based VS (LBVS), the data shown are also molecules per day and the same type of CPU

[a] ACE, COMT, AChE, and Thr have already been defined in Table 1. *PDE5* phosphodiesterase 5, *PARP* poly(ADP-ribose) polymerase

[b] Average Number of Heavy Atoms in the ligand set for each target

[c] Average number of conformations in the ligand set for each target

[d] Binding Site Grid Points

[e] Fingerprints generation

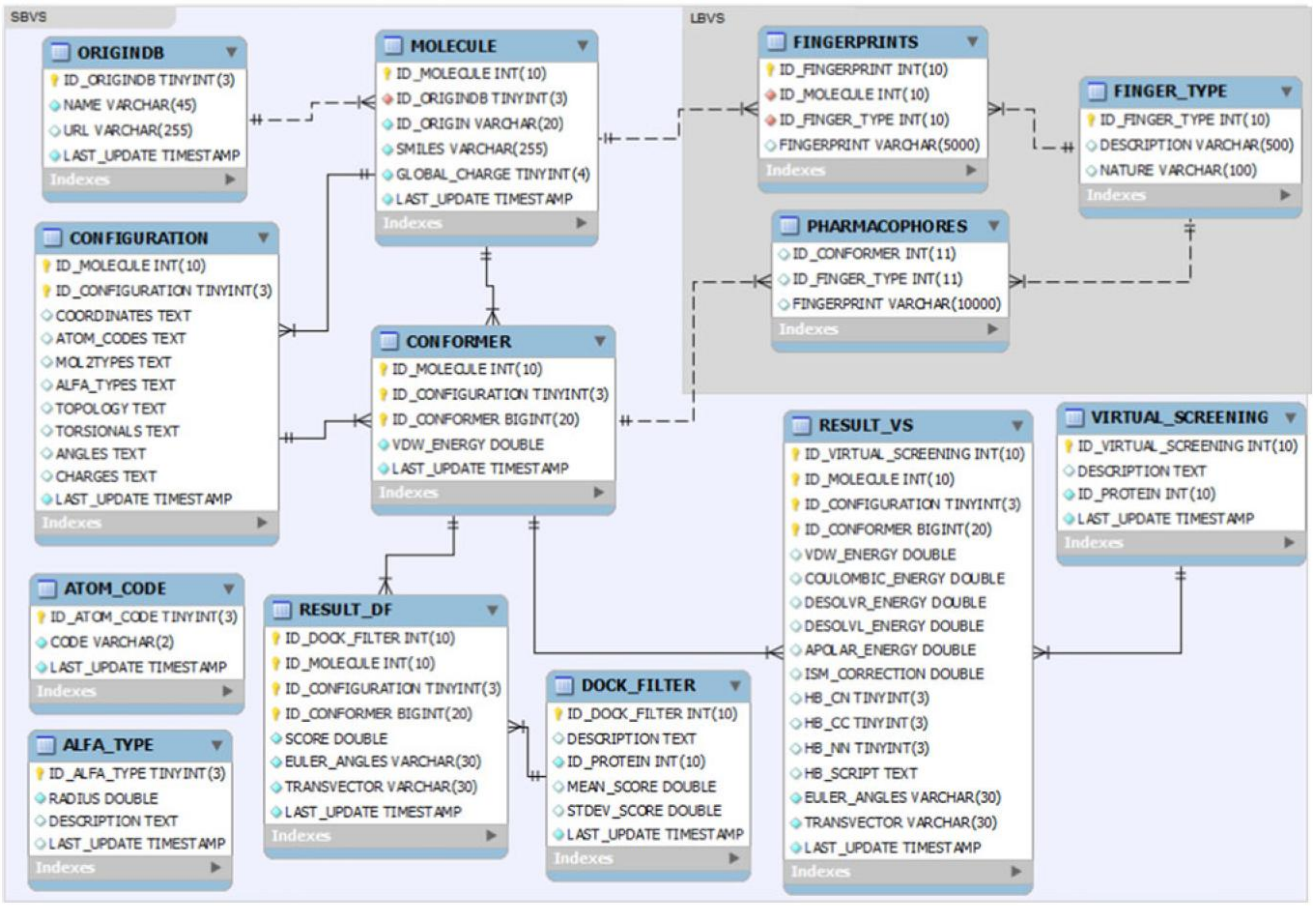


Fig. 4 Entity-relationship scheme showing the database (VSDB) used by VSDMIP. The *dark grey* shaded region corresponds to the newly added capabilities

server modules are only available for Linux with the OpenPBS/Torque queue system 2.x or above. In addition, the VS modules require an MySQL database engine and the client libraries installed on the system (mysql-server, mysql++ [34], and libmysqlclient).

The GUI front end uses PyMOL software version 1.2 or higher and requires the mysql-python (version $\geq$ 2.4) and NumPy (version $\geq$ 1.3) modules. It can be executed on Windows and Linux operating systems. The back end can be used in a wide variety of hardware architectures, from single personal computers or laptops to a cluster of processors or grid-like systems. The minimum recommended amount of main memory per processor is 1 GB. The database needs $\sim$45 GB of hard disk space to store 4×10^6 molecules together with their properties and calculated conformers. In addition, $\sim$850 and $\sim$285 MB more are needed to store information related to 2D or 3D fingerprint types, respectively. Of note, all this information is entered only once and can be reutilized in every project. Finally, $\sim$750 MB of extra storage space would be needed for the outcome of filtering/screening the entire database.

VSDMIP in the context of existing VS platforms

When VSDMIP was originally released, a limited number of similar platforms were available, namely Pipeline Pilot [7, 8] from Accelrys, alternative implementations from Schrödinger [5] and Tripos [6], and a proprietary web-based platform from Astex Therapeutics [35]. Other non-commercial solutions include the Data Management System for Distributed Virtual Screening (DVSDMS) [14], the SOMA workflow multiplatform [36], the KNIME modular environment [37], and the public access web-based DOCK Blaster platform [12]. On the other hand, the MoStBioDat database [38] was designed to store and manipulate ligand and receptor data and allows LBVS to be performed. A much wider view of the drug design cycle in terms of the implemented features is represented by OSIRIS [39] and PFAKT [40]. The former is defined as a *drug discovery informatics system* [...] *and contains a complete coverage of the drug discovery process by custom tailored applications* whereas the latter *is a suite of integrated services* [...] *that facilitate the medicinal chemistry design cycle* [...] *and provides a unified data analysis and collaboration environment.*

Some other applications have been launched with a different goal in mind, namely the need to lower the technical skill barriers so that a wider range of researchers can benefit from them. Examples are BDT [41], DOVIS [42, 43], VSDocker [44], AMMOS [45], iGEMDOCK [46], VSM-G [13], and a PyMOL plugin for AutoDock/Vina [10]. VSDMIP 1.5 represents an attempt to provide the scientific community with a customizable and comprehensive VS platform that is managed from a friendly GUI and also integrates an underlying database.

We believe, therefore, that the capabilities added to VSDMIP (namely the GUI, the LBVS module, and the interconnectivity between LBVS and SBVS modules) place version 1.5 within the *state-of-the-art* automatic platforms that perform VS experiments. Having all the tools integrated in a single application facilitates the complex task of building VS protocols and analyzing the results (Fig. 5). Besides, VSDMIP 1.5 allows the user to work with the programs individually (single docking and grid visualization, database searches, analysis of interactions, conformational analysis…) taking full advantage of the easy-to-use PyMOL interface. From the technical point of view, and although some computer skills are still required to properly configure the application for high performing computer architectures, the GUI and the configuration files provided as a guide (plus the support given by the development team) makes VSDMIP installation and maintenance relatively straightforward. VSDMIP can also be run on a desktop computer with or without the database environment, employing all the available cores in a small cluster, a user-defined number of them, or even just one. VSDMIP has been fully tested on a Linux cluster using a Linux- or Windows-running computer as the interface to the cluster. Modularity (individual tasks can be connected in different ways to allow the user to customize his/her VS protocols) and flexibility (other software pieces can be easily added to the platform and configured through user-configurable extensible markup language [XML] files) are still retained in this new version as they are considered the main cornerstones that differentiate VSDMIP from other VS platforms described to date. Detailed information related to installation, configuration, possible extensions, as well as examples, can be found within the User's Guide at http://ub.cbm.uam.es/software/vsdmip/doc/.

VSDMIP performance

The numerical results shown here for a reduced subset of targets from the DUD database follow the trend commented above [47], i.e., in general, LBVS (fingerprints) outperforms SBVS (pharmacophores and docking). On the other hand, our in-house CGRID/CDOCK docking tool ($\langle AUC \rangle \approx 0.6$) performs as well as DOCK, FlexX, ICM, and PhDOCK, and slightly worse than GLIDE and Surflex ($\langle AUC \rangle \approx 0.7$), or eHiTS ($\langle AUC \rangle \approx 0.9$) [48]. Thus, there is clearly room for improvement. Although we found these results reasonable, we are aware of the possible *analogue bias* that might be introduced during the construction of the DUD database, which leads to artificial enrichments (see Table 1) in the case of LBVS methods [49]. A key ingredient to achieve success in retrospective VS experiments is to count with a well defined database (actives + decoys) of complexes with known 3D structures and information about their activity. The term "well defined" refers to the fact that a good VS method should differentiate actives from decoys on the basis of interaction features only. The molecular properties of the selected decoys should resemble those of the active ligands and at the same time, to avoid artificial enrichments, they should

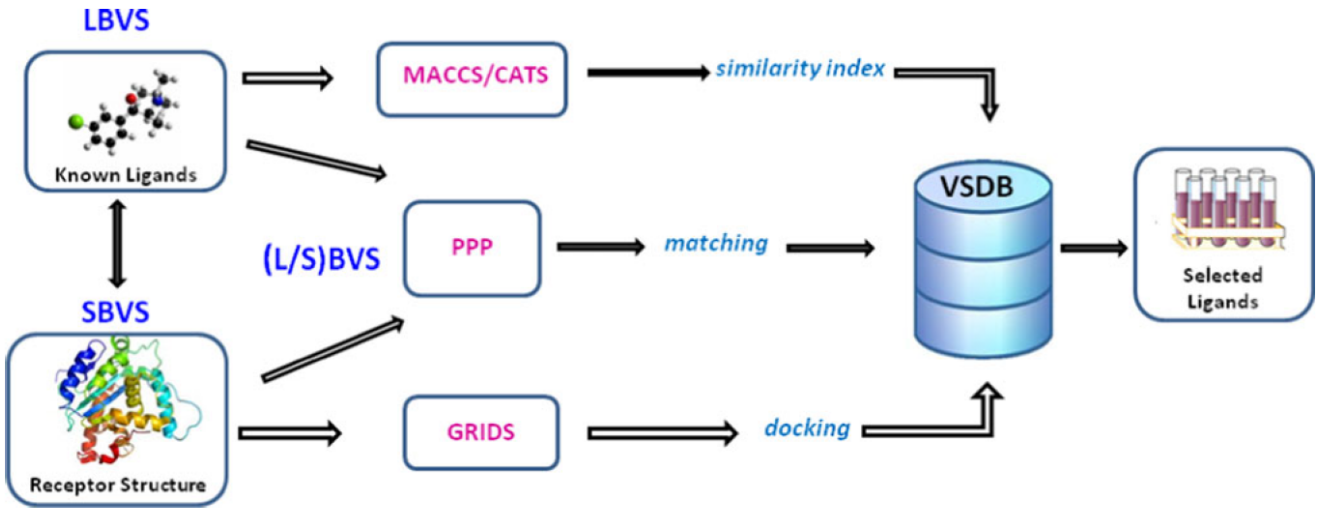


Fig. 5 VSDMIP flowchart

be structurally different. To comply with this requirement the DUD database was created as a specific benchmark to test docking methods, and has been widely accepted within the docking and VS communities to the extent that it is generally considered as the reference database. However, some caveats have been pointed out [50], in particular overfitting effects, its incomplete sampling of chemical space, and its inadequacy to be applied to LBVS methods. In relation to its applicability in LBVS, as reported here, recent results have shown that the DUD decoy sets are robust enough for LBVS [51]. Furthermore, it was found that 2D fingerprints outperformed 3D shape-based approaches as VS tools, in agreement with a previous study where fingerprints, pharmacophores, and docking were compared as VS engines [47]. This behavior might be linked to the *analogue bias* effect, so called because chemical diversity is not warranted in the DUD database and many decoys share a reduced set of common scaffolds resembling the actives. If by any chance one of the most populated scaffolds is selected as the query, artificial enrichment comes into play. *Analogue bias* is more prone to occur with the use of fingerprints than with 3D methods based on shape, pharmacophores or docking simply because the former represent molecular similarity whereas the latter involve chemical interactions. This observation was confirmed by us in some of the targets. We calculated similarity matrices among all the actives for each target using the *Tc* (see Eq. 1), coloring them from blue (low similarity, $Tc = 0$) to red (high similarity, $Tc = 1$). Selected results are shown in Fig. 3. For HMGR and TK the active sets are very similar (predominant red color in both matrices) with MACCS AUCs $\sim$ 0.8–0.9, CATS AUCs $\sim$ 0.8, and CDOCK AUCs $\sim$ 0.6. On the other hand, more difficult cases are represented by COX1 and ALR2, where the active sets are highly dissimilar, as shown by the prevailing blue color. In these cases, MACCS and CATS AUCs drop to 0.5–0.6, while CDOCK AUCs are less sensitive and only small variations are observed.

Benchmarking

With a medium-size cluster ($\approx$100 processors) and a database of the order of several million compounds (4×10^6 in VSDMIP), it would take 3.6 months to insert all the molecules. Although this can be considered a very long time, most of it is spent on calculating the point charges, and we recall here that the molecules need to be inserted only once and then can be reutilized as many times as desired. Docking the complete database would take around 5 months. Nevertheless, and given the high speed obtained in LBVS, several of these runs can be easily completed in an affordable time span before undertaking more computationally intensive SBVS. LBVS would then serve as a filter, leading to an important reduction in the number of molecules to be docked, as shown above for the ACE and MR targets, and as a result, to an important optimization of the computer resources. By the time the results presented here were collected, we acquired some new Xeon Core2 64-bit 2.5 GHz processors, and preliminary tests yielded a speed-up of 2.2-fold relative to the old ones. Therefore, it should now be possible to insert 4×10^6 SMILES strings in less than 2 months, and dock the entire database in 2.2 months using 100 processors of this type. In view of these figures, and taking into account the ever-increasing computer power and the fact that docking (in the way it has been implemented in VSDMIP) is 100% scalable (the more processors available, the less time required to complete the tasks), we believe that VSDMIP is a computational platform capable of performing VS experiments in perfectly affordable time schedules in an environment accessible to a large number of researchers.

Conclusions

VSDMIP 1.5 allows an inexperienced user to execute both SBVS and LBVS protocols, or any combination of the two, by means of an easy-to-learn and friendly GUI implemented in the commonly used PyMOL molecular graphics program. We have tested its ability to conduct VS protocols and compared the efficiency of different methods. Good agreement with results from previous studies was found but we also realized that the *analogue bias* effect in the DUD database can lead to artificial enrichment for LBVS. In terms of computer time, we show that VSDMIP can indeed cope with the current demand of performing VS experiments in weeks rather than in months. This version of the platform is distributed to the scientific community upon request from the authors as a bundled package including the scripts and necessary SQL files to create the database structure and the XML configuration files. The programs implemented in the platform (except those that need to be purchased for a modest prize, such as CORINA or AMBER) are either free for academics (MOPAC, DOCK, FRED, AutoDock) or will be released under a scientific/academic non-profit and non-commercial license as is the case for ALFA, CGRID, CDOCK, and ISM.

Acknowledgments The authors thank Dr. Eva M[a] Priego and Dr. Alberto Gómez for testing the application and valuable comments, as well as the rest of members of the Bioinformatics Unit at CBMSO and the Molecular Modeling group at UAH for encouragement and fruitful discussions. This work was supported by grants from Ministerio de Ciencia e Innovación (MICINN) BIO2008-04384 (to A. M.) and SAF2009-13914-C02-02 (to F. G.), and Comunidad Autónoma de

Madrid (CAM) S-BIO-0214-2006. A. M. acknowledges CAM for financial support through the AMAROUTO program to the Fundación Severo Ochoa, R. G. -R. thanks MICINN for a contract from "Programa de Personal Técnico y de Apoyo 2008", and A. C. thanks Ministerio de Educación for the FPU grant AP2009-0203. We are grateful to OpenEye Scientific Software, Inc. for providing us with an academic license for their software. The technical support and advice from the Bioinformatics Facility at CBMSO is gratefully acknowledged, as well as the computer resources, technical expertise and assistance provided by the Barcelona Supercomputing Center—Centro Nacional de Supercomputación.

References

1. Munos B (2009) Nat Rev Drug Discov 8(12):959
2. Jorgensen WL (2004) Science 303(5665):1813
3. Zhou HX, Gilson MK (2009) Chem Rev 109(9):4092
4. Ivanov AS, Veselovsky AV, Dubanov AV, Skvortsov VS (2006) Methods Mol Biol 316:389
5. Maestro (2011) Maestro, version 9.2. Schrödinger, LLC, New York
6. SYBYL-X 1.2. (2011) Tripos International, 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA
7. Schulz T, Pleiss J, Schmid R (2000) Protein Sci 9(6):1053
8. Hassan M, Brown RD, Varma-O'brien S, Rogers D (2006) Mol Divers 10(3):283
9. DeLano WL (2002) The PyMOL molecular graphics system. Schrodinger Inc, New York
10. Seeliger D, de Groot BL (2010) J Comput Aided Mol Des 24(5):417
11. Lill MA, Danielson ML (2011) J Comput Aided Mol Des 25(1):13
12. Irwin JJ, Shoichet BK, Mysinger MM, Huang N, Colizzi F, Wassam P, Cao Y (2009) J Med Chem 52(18):5712
13. Beautrait A, Leroux V, Chavent M, Ghemtio L, Devignes MD, Smail-Tabbone M, Cai W, Shao X, Moreau G, Bladon P, Yao J, Maigret B (2008) J Mol Model 14(2):135
14. Zhou T, Caflisch A (2009) J Chem Inf Model 49(1):145
15. Gil-Redondo R, Estrada J, Morreale A, Herranz F, Sancho J, Ortiz AR (2009) J Comput Aided Mol Des 23(3):171
16. Huang N, Shoichet BK, Irwin JJ (2006) J Med Chem 49(23):6789
17. Murray CW, Baxter CA, Frenkel AD (1999) J Comput Aided Mol Des 13(6):547
18. Schneider G, Neidhart W, Giller T, Schmid G (1999) Angew Chem Int Ed 38(19):2894
19. Open Babel: The open source chemistry toolbox; 2011
20. Weininger D (1988) J Chem Inf Comput Sci 28(1):31
21. Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ (2005) J Comput Chem 26(16):1668
22. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) J Chem Phys 79:926
23. Zacharias M, Luty BA, Davis ME, McCammon JA (1994) J Mol Biol 238(3):455
24. Pastor M, Cruciani G (1995) J Med Chem 38(23):4637
25. Perez C, Ortiz AR (2001) J Med Chem 44(23):3768
26. Weininger D (1988) J Chem Inf Model 28(1):31
27. Corina. Computerchemie Langemarckplatz 1, Erlangen, Germany: Molecular Networks GmbH; 2000
28. Stewart JJ (1990) J Comput Aided Mol Des 4(1):1
29. Gil Redondo R (2006). Master thesis. UNED, Madrid
30. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA (2004) Nucleic Acids Res 32(suppl 2):W665
31. Li H, Robertson AD, Jensen JH (2005) Protein Struct Funct Bioinforma 61(4):704
32. Triballeau N, Acher F, Brabet I, Pin JP, Bertrand HO (2005) J Med Chem 48(7):2534
33. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner J, Egon L, Willighagen O (2006) J chem inf model 46(3):991
34. MySQL++. A MySQL API for C++: Tangensoft
35. Watson P, Verdonk M, Hartshorn MJ (2003) J Mol Graph Model 22(1):71
36. Lehtovuori PT, Nyronen TH (2006) J Chem Inf Model 46(2):620
37. Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R, Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B (2008) KNIME The Konstanz information miner. In: Bock HH, Gaul W, Schader M et al (eds) Data analysis. machine learning and applications. Springer, Berlin Heidelberg, p 319
38. Bak A, Polanski J, Kurczyk A (2009) Molecules 14(9):3436
39. Sander T, Freyss J, von Korff M, Reich JR, Rufener C (2009) J Chem Inf Model 49(2):232–246
40. Brodney MD, Brosius AD, Gregory T, Heck SD, Klug-McLeod JL, Poss CS (2009) J Chem Inf Model 49(12):2639
41. Vaque M, Arola A, Aliagas C, Pujadas G (2006) Bioinformatics 22(14):1803
42. Zhang S, Kumar K, Jiang X, Wallqvist A, Reifman J (2008) BMC Bioinformatics 9:126
43. Jiang X, Kumar K, Hu X, Wallqvist A, Reifman J (2008) Chem Cent J 2:18
44. Prakhov ND, Chernorudskiy AL, Gainullin MR (2010) Bioinformatics 26(10):1374
45. Pencheva T, Lagorce D, Pajeva I, Villoutreix BO, Miteva MA (2008) BMC Bioinformatics 9:438
46. Hsu KC, Chen YF, Lin SR, Yang JM (2011) BMC Bioinformatics 12(Suppl 1):S33
47. Modest von Korff JF, Sander Thomas (2009) J Chem Inf Model 49(2):209
48. Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, Hu Y, Humblet C (2009) J Chem Inf Model 49(6):1455
49. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WT, Murray CW, Taylor RD, Watson P (2004) J Chem Inf Comput Sci 44(3):793
50. Irwin JJ (2008) J Comput Aided Mol Des 22(3–4):193
51. Venkatraman V, Perez-Nueno VI, Mavridis L, Ritchie DW (2010) J Chem Inf Model 50(12):2079

Article II

CRDOCK: an ultrafast multipurpose protein-ligand docking tool

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

CRDOCK: An Ultrafast Multipurpose Protein−Ligand Docking Tool

Álvaro Cortés Cabrera,[†,‡] Javier Klett,[‡] Helena G. Dos Santos,[‡] Almudena Perona,[‡,§] Rubén Gil-Redondo,[‡,§] Sandrea M. Francis,[∥] Eva M. Priego,[⊥] Federico Gago,[†] and Antonio Morreale*[,‡]

[†]Departamento de Farmacología, Universidad de Alcalá, E-28871 Alcalá de Henares, Madrid, Spain

[‡]Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Campus UAM, c/Nicolás Cabrera 1, E-28049 Madrid, Spain

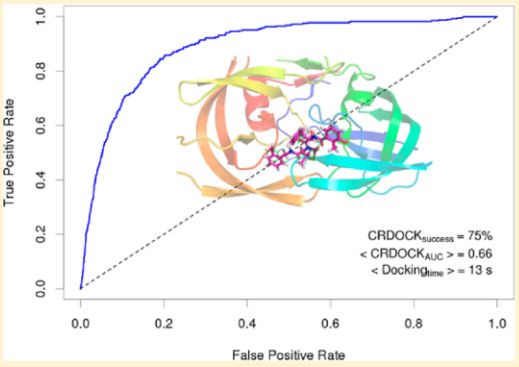
[§]SmartLigs Bioinformática S.L., Fundación Parque Científico de Madrid, c/Faraday 7, Campus de Cantoblanco UAM, E-28049 Madrid, Spain

[∥]Instituto de Biomedicina de Valencia (IBV-CSIC), c/Jaime Roig 11, E-46010 Valencia, Spain

[⊥]Instituto de Química Médica (CSIC), c/Juan de la Cierva 3, E-28006 Madrid, Spain

Supporting Information

ABSTRACT: An ultrafast docking and virtual screening program, CRDOCK, is presented that contains (1) a search engine that can use a variety of sampling methods and an initial energy evaluation function, (2) several energy minimization algorithms for fine tuning the binding poses, and (3) different scoring functions. This modularity ensures the easy configuration of custom-made protocols that can be optimized depending on the problem in hand. CRDOCK employs a precomputed library of ligand conformations that are initially generated from one-dimensional SMILES strings. Testing CRDOCK on two widely used benchmarks, the ASTEX diverse set and the Directory of Useful Decoys, yielded a success rate of ∼75% in pose prediction and an average AUC of 0.66. A typical ligand can be docked, on average, in just ∼13 s. Extension to a representative group of pharmacologically relevant G protein-coupled receptors that have been recently cocrystallized with some selective ligands allowed us to demonstrate the utility of this tool and also highlight some current limitations. CRDOCK is now included within VSDMIP, our integrated platform for drug discovery.



1. INTRODUCTION

Docking and virtual screening (VS) strategies have acquired a relevant role in modern drug discovery since the pioneering work of Kuntz et al.[1] back in the early 1980s. However, and despite many advances carried out in the field during the past decade, this methodology is still far from perfect.[2] To increase its usefulness, more accurate methods are needed that can not only predict the native pose of a ligand within a protein in a crystallographic structure at the top of the list of possible solutions, as done in docking studies, but also discriminate true binders from a pool of decoys, as done in VS. Moreover, a modern docking tool needs to be fast because the number of molecules in currently used chemical libraries is well above 10^6. It is clear then that different objectives are pursued in docking and VS. In the former, native pose prediction is the main goal; in the latter, as long as true binders are separated from decoys, less accurate binding poses can be tolerated. In an ideal case, however, both criteria should be met because success in VS for the wrong reasons is unlikely to be reproducible and does not contribute to advancing the field.[3] More commonly, the goodness of fit between the ligand and the receptor is evaluated by means of an energy function composed of different terms that attempt to account for the forces driving the binding event. Although the underlying physical laws describing the binding process are well understood, accuracy and computational resources (mainly time) evolve in opposite directions, and fine tuning the appropriate balance between them is by no means an easy task. Therefore, accuracy is normally sacrificed for speed, especially in VS, and very often too simplistic scoring functions are employed.

Prior to scoring it is also necessary to sample the binding site of the receptor as exhaustively as possible. To this end and to save computer time, the space is usually discretized on a three-dimensional (3D) lattice and probe interaction energies at the grid points are calculated and stored.[4] Then, the molecule under study is translated and rotated at each lattice node along the three dimensions of the box, and interaction energies with the protein are estimated for each pose using the data stored on the grid points. Depending on the docking tool, the conformers can be generated within the binding site itself "on the fly" or created beforehand and stored to be reused again as many times as needed. The former method is more computer intensive, but as an advantage, it can generate strained

Received: April 19, 2012
Published: July 5, 2012

dx.doi.org/10.1021/ci300194a | J. Chem. Inf. Model. 2012, 52, 2300−2309

conformations that adapt better to the active site environment. If the latter method is employed, a collection of all allowed conformers is quickly generated only once following some predefined rules, but the drawback is that a relevant conformation can be missed. For example, AutoDock[5] and GOLD[6] produce conformers in situ, whereas Glide[7,8] and FRED[9] use a pregenerated database of conformers. Our original in-house docking tool, CDOCK,[10] belongs to this second category because it was developed bearing in mind its potential use in VS, where millions of small molecules, some of them with hundreds of possible conformers, are available for docking in different projects. CDOCK was implemented in our open-access Virtual Screening Data Management on an Integrated Platform (VSDMIP) which has been recently extended to cover not only receptor-based VS[11] but also ligand-based[12] and fragment-based VS.[13]

Among the challenges still to be faced, some are more technical and related to computational times and correct implementation of tools to configure distinct VS protocols while others have to do with explicit inclusion of entropy,[14] solvent effects,[15,16] and receptor flexibility,[17,18] which appear to be necessary to get more accurate estimates of binding free energies. Indeed, theoretical approaches that tackle these problems continue to be developed and improved but are usually impractical for large VS campaigns because of the huge number of compounds that need to be evaluated. These computational efforts can be highly reduced if a prioritized list of compounds is available to be passed on to the most demanding calculations. This should be, in fact, the final objective of a docking program: saving time without increasing the false positive and negative rates compared to random selection. Unfortunately, not many currently available tools meet this requirement.

Bearing these facts in mind and in an attempt to improve the sampling and scoring capabilities of CDOCK, we present here CRDOCK, an ultrafast ligand docking program that was tested on two widely used benchmarks: the ASTEX diverse set (ADS, for docking)[19] and the Directory of Useful Decoys (DUD, for VS).[20] The latter was subsequently expanded with a representative group of recently available and pharmacologically very relevant G protein-coupled receptors (GPCR) in complex with some selective ligands. Being aware of the importance of water-mediated interactions in receptor−ligand binding, a water selection algorithm was implemented that is based on interaction energy calculations on a 3D grid, as pioneered by Peter Goodford in his renowned GRID program.[4] We also detected, discussed, and in most cases solved some widely reported problems[19,21] involving several well-known ligand−receptor complexes.

2. METHODS

Our CRDOCK tool contains (1) a search engine that can use a variety of sampling algorithms and an initial energy evaluation function for placing the ligand in the binding site, (2) several energy minimization algorithms for fine tuning the binding poses, and (3) different scoring functions for ligand ranking. Different methods are available for each of these components, and they can be independently chosen and combined.

2.1. Ligand Preparation. The ligands present in the complexes studied were prepared in two different ways. To test the docking engine alone, their X-ray coordinates from the PDB files were extracted and the ligand-free protein was used as the target (rigid ligand docking). Protonation and tautomeric states for these ligands were assigned with Open Babel 2.3.0[22] assuming a pH of 7.0, and no further manipulation of the coordinates was performed so as to preserve the native bound conformations. As a more realistic alternative and to check the efficiency of our in-house conformer generator, we also started from scratch, in the absence of any information about the ligand's 3D structure (flexible ligand docking). To this end, each ligand from the previous step was converted into a 1D simplified molecular-input line-entry system (SMILES) string and then inserted into the VSDMIP database following our standard protocol: (a) automated conversion from SMILES to 3D MOL2 format using CORINA,[23] (b) atomic charge calculations with MOPAC[24] (AM1 ESP method) on every single structure provided by CORINA, (c) atom-type assignment according to the AMBER force field,[25] and (d) conformer generation using ALFA.[26]

2.2. CRDOCK Constituent Parts. The three main components of the CRDOCK tool are depicted in Figure 1.

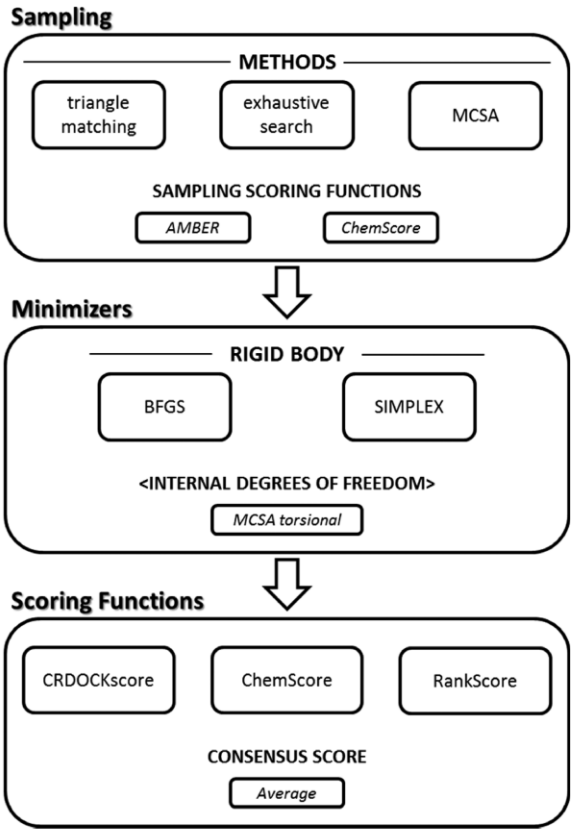


Figure 1. Graphical overview of CRDOCK components and workflow. Method enclosed in angle brackets is optional.

The different methods integrated in each step can be independently selected and later combined to configure different custom-made workflows. This is of special interest in those cases where a researcher faces computational restrictions.

2.2.1. Sampling. CRDOCK implements three sampling strategies: (a) triangle matching, (b) exhaustive search, and (c) Monte Carlo simulated annealing (MCSA).

a. Triangle Matching. CRDOCK determines the interaction points for a given ligand conformation using the functional groups present in its structure. All possible combinations of 3 different interaction points (for each conformer) are generated

applying a cutoff to avoid very short edges (4.5 Å by default). These are the ligand interaction triangles. The same combination is performed for the receptor interaction points (see receptor active site analysis below) to obtain the receptor interaction triangles. The program then looks for the best superimpositions between the receptor's and the ligand's interaction triangles, evaluating each one with an AMBER-like (12−6 Lennard−Jones potential with an electrostatic term modeled with a sigmoidal dielectric screening function) or the ChemScore[27] empirical scoring function. If none of the ligand conformers can be used to build triangles, a lower cutoff for the edges is employed (2.5 Å). Finally, if no triangle can be built with this reduced distance, the exhaustive search or MCSA algorithms (see below) will be used instead.

b. Exhaustive Systematic Search. In this case each ligand conformer is translated over each single grid point and rotated in all directions with steps of 30° on each axis. The interaction energy for each generated pose is evaluated with the energy functions described in point a above. Because this search is quite time consuming, it is performed only if the number of conformers is ≤5. For the remaining cases the MCSA method described below is employed.

c. Monte Carlo Simulated Annealing. Random translations and rotations of the ligand are generated to determine a new pose from the last accepted pose (the first pose is generated randomly). The new pose is accepted or rejected depending on a probability managed by the temperature (Metropolis criterion). By default and on the basis of results from internal tests, 23 temperature rounds are performed with a maximum number of generated poses per round of 725 000 to avoid redundancy. The algorithm starts at 773 K, and this temperature is scaled down at each round by 20%. The probability of change for each parameter defining a pose was set to 0.8.

Regardless of the sampling method, the best 512 poses per conformer are saved in a stack up to a maximum of 5000 per molecule. The final result is a collection of the best ligand poses (5000 by default) which can be passed on to the next step to be refined by energy minimization. In order to promote diversity, it is possible to apply a root-mean-square deviation (rmsd)-based filter to check whether or not a new pose is added to the list.

2.2.2. Minimization. Prior to scoring, the above selected poses can be refined either by rigid-body energy minimization (translations and rotations) or by changing the internal degrees of freedom (torsional angles) to fine tune their fit within the receptor binding site. This process is carried out expediently using precalculated AMBER-like potential interaction energies stored in a 3D grid.[28,29] Three algorithms have been implemented: (a) Broyden−Fletcher−Goldfarb−Shanno (BFGS, for rigid-body minimizations), a deterministic method that belongs to the family of quasi-Newton methods; (b) Amoeba or downhill SIMPLEX (for rigid-body minimizations), the stochastic algorithm from Nelder and Mead;[30] and (c) MCSA torsional (MCSAtor, for internal degrees of freedom [torsions] minimizations), a stochastic method analogous to the described MCSA algorithm but that only optimizes molecular torsions. By default, 200 steps of BFGS rigid-body minimization are performed.

2.2.3. Scoring Functions. The final ranking of the resulting poses from the previous step can be evaluated with either a single scoring function or a combination of several of them ("consensus scoring"). The available options are (a) CRDOCKscore, (b) ChemScore, and (c) RankScore.

a. CRDOCKscore. This is a modified version of GlideScore,[31] the scoring function implemented in program Glide (eq 1), that combines van der Waals (E_{vdW}) and electrostatic (E_{qq}) energy terms from the AMBER force field with lipophilic (E_{lipo}) and hydrogen-bonding (E_{hb}) terms from ChemScore.[27] The AMBER terms are scaled by weighting factors α and β for E_{vdW} and E_{qq}, respectively, as defined in GlideScore.

$$\text{CRDOCKscore} = \alpha E_{vdw} + \beta E_{qq} + E_{lipo} + E_{hb} \quad (1)$$

where α = 0.065 and β = 0.130. No additional modifications were made to the AMBER or ChemScore lipophilic and hydrogen-bonding terms.

b. ChemScore. ChemScore is our implementation of the well-known ChemScore[27] scoring function.

c. RankScore. RankScore is a statistical potential scoring function specifically derived for VS classification from known sets of compounds (DUD).[32]

2.2.4. Water Selection Algorithm. Very often water-mediated interactions between the ligand and the receptor binding site are key for accurate fitting. We employed a modified version of cGRILL (see receptor active site analysis) to generate water affinity maps using a probe representing a water molecule that can act as a hydrogen-bond acceptor and donor. For this purpose we employed the concept of *extended atom* using an AMBER-like energy function for an oxygen atom endowed with a van der Waals term, a partial charge of −0.12 au and a hydrogen-bond block function based on an ideal H-acceptor distance of 1.8 Å and a donor−H-acceptor angle of 180°. Then the program clusters similar interaction areas using an energy cutoff which is one-half of the maximum value of the scoring function.

2.3. Benchmarks. Three different test sets were used: (a) ADS for pose prediction, (b) DUD for VS, and (c) GPCR for pose prediction and VS. The standard criterion to validate the ability to predict the native pose was the heavy atoms rmsd between the docking solution and the native conformation for each ligand in the crystal structure. In common with other similar studies, we chose an rmsd value of 2 Å as the upper limit for a solution to be considered correct. VS performance was assessed by means of the area under the curve (AUC) of the generated receiver operating characteristics (ROC) plots.[33]

All receptors (except otherwise stated) were prepared for both docking and VS following the same protocol, namely, for each one, all species other than the protein itself, cofactors, and metals were removed. Hydrogen atoms were added using the pdb2pqr[34] tool and adapted to the AMBER 03 force field using GROMACS v.4.5.3. One thousand steps of steepest descent were followed by 2000 steps of Polak−Ribiere conjugate gradient energy minimization where only hydrogen atoms were allowed to move. No energy minimization was performed for DUD targets so that our results could be compared with those published in other DUD-related publications. The cubic grid for cGRILL calculations was defined as the space delimited by the axis-parallel box containing the cocrystallized ligand, augmented by 5 Å in each axis direction.

a. ASTEX Diverse Set (ADS). The ADS is composed of 85 protein−ligand complexes that can be downloaded from the Protein Data Bank (PDB).

b. Directory of Useful Decoys (DUD). The DUD is composed of 40 different targets with known 3D structures and a set of true/fake binders for each target.

c. G Protein-Coupled Receptors (GPCR) Set. We enlarged the original DUD by including such pharmacologically relevant targets from the GPCR family as the adrenergic β_2, the dopaminergic D_3, the muscarinic M_2, the histamine H_1, and the opioid μ receptors, all of which have been cocrystallized in the presence of antagonists or inverse agonists at a resolution ≤ 3.1 Å. For both self-docking and VS tests we used the ligand-bound protein structures found in PDB entries 2RH1 (β_2), 3PBL (D_3), 3UON (M_2), 3RZE (H_1), and 4DKL (μ). In the latter case, the covalent bond between the morphinan ligand and Lys233 was broken and an alternate nonclashing rotamer was selected for Lys233 using PyMOL.[35] SMILES strings for known true ligands for these targets (Supporting Information, Table S1) were taken from the DrugBank database.[36] To select a suitable set of decoys, we followed a similar procedure to that reported in the original DUD description: (a) the clean drug-like set of small molecules was downloaded from ZINC[37] (9 542 593 SMILES strings); (b) Molecular ACCess System (MACCS) fingerprints were calculated for each SMILES string using OpenBabel;[22] (c) a Tanimoto cutoff of 0.4 was used as a filter to select the most topologically dissimilar compounds as compared to the known ligands (415 636); (d) Qikprop[38] was used to calculate physicochemical properties for each compound and select those most similar to the true ligands; and (e) 30 decoys per ligand were saved for the VS experiments.

2.4. Receptor Ligand-Binding Pocket Analysis. Given a suitably prepared receptor structure, the next steps are to energetically characterize the active site and to determine its main interaction points.

a. Energetic Characterization. This is performed with the cGRILL program, an improved version of our CGRID code[10] that relies on interaction energy calculations on 3D grids as pioneered by Goodford in his well-known GRID program.[4] cGRILL uses the AMBER 12−6 Lennard−Jones term for C, N O, H, S, and P probe atoms and an electrostatic term modeled with a sigmoidal dielectric screening function, together with other terms from ChemScore (lipophilic, H-bond acceptor, H-bond donor, "mixture", metal, and clash). Finally, a grid containing clash-free points was added for accelerating the sampling process during docking. All grid maps can be inspected using PyMOL molecular visualization software.[35] Residues, cofactors, and other species are parametrized automatically using an AMBER force field-like atom-typing scheme.

b. Interaction Points. The best interaction areas ("hot spots") are mapped by sampling the active site with different molecular probes, an idea already introduced in other docking programs.[39] The calculations are similar to those in cGRILL for AMBER grids but using polyatomic probes instead, which are allowed to rotate at each grid point with a step size of 180° in each axis. The molecular probes are CH_4 to detect lipophilic regions and NH and CO to detect H-bond-accepting and -donating partners, respectively. The best result for any of the probes at each grid point is selected. The generated set is postprocessed to avoid redundancy. For a given area, only the best probe in a radius of 2 Å for lipophilic or 1.5 Å for hydrogen-bond probes is kept. Then, these nonredundant probes are rescored by summing up their neighbors' energetic scores.[39] The best probe with the highest new score is selected as the best point in the active site, and all the surrounding points with a distance less than 4 Å are also selected. The process continues adding points within 4 Å from the selected ones until no more points fulfilling the distance restraints are available. The selected docked probes are exported as a PDB file.

2.5. Alternative Docking Protocols. Two different alternatives were explored.

(a) Alternative 1. Our original CDOCK code with default parameters: MCSA/exhaustive sampling, SIMPLEX refinement, and the scoring function as the sum of van der Waals and electrostatic terms from the AMBER force field, the electrostatic contributions to ligand and receptor desolvation, the nonelectrostatic part of the desolvation modeled as a linear relationship with the solvent accessible surface area lost once the complex has been formed, and an explicit term to account for hydrogen-bonding interactions. In the preparation of the receptor, water molecules involved in the binding event are always kept and full complex relaxation is accomplished using energy minimization.

(b) Alternative 2. A different CRDOCK configuration consisting of (1) default sampling (triangle matching or MCSA), (2) BFGS and MCSAtor for fine tuning, and (3) nonbonding energy terms (E_{vdW} and E_{qq}) from the AMBER force field for pose scoring. For the VS experiments, ChemScore, AMBER, and RankScore were used as the scoring functions as well as a simple average of the three ("consensus scoring").

3. RESULTS AND DISCUSSION

Different CRDOCK configurations were tested for both data sets. After analyzing the results (Table 1), the combination that yielded the best performance was triangle matching and AMBER energy evaluation, BFGS as the rigid body minimizer without torsional optimization, and the CRDOCK scoring function. The results shown in Table 2 were obtained with this CRDOCK configuration.

Table 1. Summary of the Results from Alternative Protocol Using the ADS and the DUD

protocol	no. complexes with rmsd ≤ 2.0 Å (ADS)	average AUC (DUD)
CRDOCK	62	0.66
alternative 1 (CDOCK)	64[e]	0.60
alternative 2 FF[a]	59[f]	0.56
alternative 2 CS[b]	59[f]	0.58
alternative 2 RS[c]	59[f]	0.57
alternative 2[d]	59[f]	0.58

[a]AMBER force field scoring function. [b]ChemScore. [c]RankScore. [d]Consensus scoring = average between AMBER, Chemscore, and Rankscore scoring functions. [e]See methods. [f]Selected poses are always the same, and the difference is due to the final scoring function applied.

3.1. Docking and Scoring: Pose Prediction and Errors Found. Although with some controversy,[40] the rmsd between a docking solution and the crystallographic coordinates is still the most widely accepted criterion as a metric of success. A docking solution with an rmsd value ≤ 2.0 Å is regarded as a correct pose. Considering the receptor as a rigid entity, two different docking experiments were conducted: (a) rigid ligand docking, where the ligand conformation is taken directly from the structure of the complex, and (b) flexible ligand docking, where

Table 2. Summary of the VS Statistics Related to the AUC Values Grouped by Protein Families

target[a] family	average	SD[b]	median	max	min
global	0.66	0.15	0.69	0.97	0.32
kinases[c]	0.63	0.11	0.63	0.77	0.47
serineproteases[d]	0.78	0.06	0.76	0.84	0.74
NHR[e]	0.71	0.19	0.74	0.97	0.42
metalloenzymes[f]	0.61	0.15	0.63	0.76	0.41
folateenzymes[g]	0.83	0.15	0.83	0.94	0.72
rest	0.61	0.26	0.59	0.80	0.32

[a]ACE, angiotensin-converting enzyme; AChE, acetylcholinesterase; ADA, adenosine deaminase; ALR2, aldose reductase; AmpC, AmpC β-lactamase; AR, androgen receptor; CDK2, cyclin-dependent kinase 2; COMT, catechol *O*-methyltransferase; COX-1, cyclooxygenase-1; COX-2, cyclooxygenase-2; DHFR, dihydrofolate reductase; EGFR, epidermal growth factor receptor; ER_{ago}, estrogen receptor (agonist-bound conformation); ER_{antago}, estrogen receptor (antagonist-bound conformation); FGFR1, fibroblast growth factor receptor kinase; FXa, factor Xa; GART, glycinamide ribonucleotide transformylase; GPβ, glycogen phosphorylase β; GR, glucocorticoid receptor; HIVPR, HIV protease; HIVRT, HIV reverse transcriptase; HMGR, hydroxymethyl-glutaryl-CoA reductase; HSP90, human heat shock protein 90; INHA, enoyl ACP reductase; MR, mineralocorticoid receptor; NA, neuraminidase; P38 MAP, P38 mitogen-activated protein; PARP, poly(ADP-ribose) polymerase; PDE5, phosphodiesterase 5; PDGFRB, platelet-derived growth factor receptor kinase; PNP, purine nucleoside phosphorylase; PPARγ, peroxisome proliferator activated receptor γ; PR, progesterone receptor; RXRα, retinoic X receptor α; SAHH, *S*-adenosyl-homocysteine hydrolase; SRC, tyrosine kinase SRC; Thr, thrombin; TK, thymidine kinase; VEGFR2, vascular endotelial growth factor. receptor; [b]Standard deviation. [c]CDK2, EGFR, FGFR1, HSP90, P38 MAP, PDGFRB, SRC, TK, and VEGFR2. [d]FXa, Thr, and trypsin. [e]AR, ER_{ago}, ER_{antago}, GR, MR, PPARγ, PR, and RXRα. [f]ACE, ADA, COMT, and PDE5. [g]DHFR and GART.

a precalculated set of conformers is generated from scratch. In the former we challenge, on one hand, whether the sampling algorithms are able to generate the correct pose and, on the other hand, whether the scoring function is able to recognize it as its best solution. In the latter we also test the conformer generator. Our results (Figure 2) indicate that CRDOCK correctly identifies 93% (79 out of 85) of the native poses in the rigid docking experiment. However, this percentage falls to 73% (62 out of 85) when the ligand coordinates are generated from the SMILES strings using CORINA and ALFA.[26] Both results are comparable to those obtained with the most widely used docking programs. For example, in the original ADS study the authors described a very similar performance for GOLD (around 70−80% of complexes with an rmsd $\leq$ 2.0 A) and almost the same decrease in performance when the ligand 3D structure was generated from scratch instead of using the crystallographic pose.[19] Li et al. used a set of 195 diverse high-resolution ligand−protein complexes to compare Glide, GOLD, LigandFit, and Surflex docking programs and obtained a variable range from 60% to 80% in most cases.[41] Cross et al., using a subset of high-resolution structures from the CCCD/Astex set, compared GLIDE, ICM, PhDock, FlexX, and Surflex docking tools. The 70−90% rate of successfully docked poses when the native ligand was reduced to 50−77% when the ligand structure was generated from scratch using CORINA.[42] More recently and using ADS, performances of 60−80% for Surflex-Dock[43] and ~74% for LeadFinder[44] were reported. All benchmarks confirm than the results currently obtained with CRDOCK are comparable to those obtained with other modern docking programs. Next, we looked for the reasons for failure.

Common Errors. Some errors that we encountered in flexible ligand docking appear to be commonly reported by others[19] using different docking engines and scoring functions.

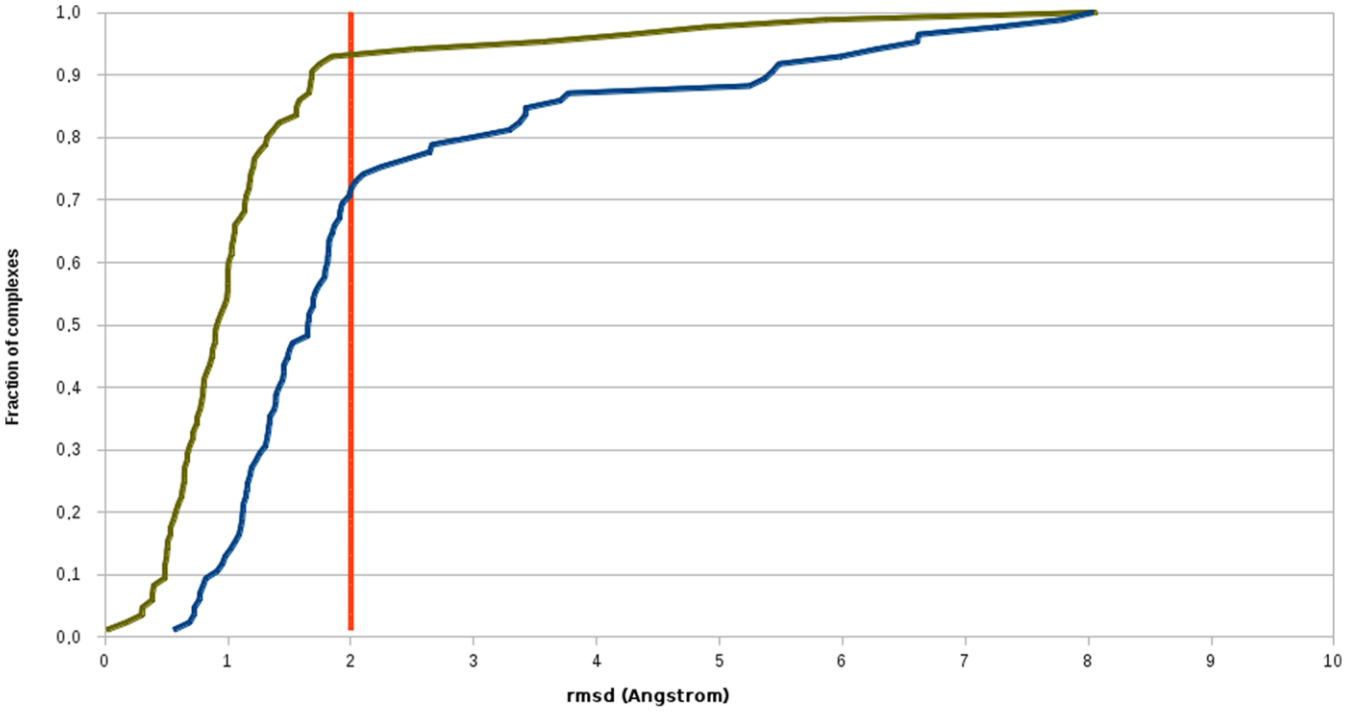


Figure 2. Pose prediction performance obtained using CRDOCK in rigid (green) and flexible (blue) ligand docking. Red line indicates a success threshold of rmsd = 2 Å.

The ligands in PDB entries 1JJE, 1SJ0, 1YVF, and 1TZ8 belong to this category (Supporting Information, Figure S1). The ligand in complex 1JJE displays an almost planar and quasi-symmetrical conformation with a central diacetal moiety chelating a zinc ion. Due to these properties it is not unreasonable that the docking program, while being able to capture the native interactions, selects a pose that is rotated by 180° compared to the X-ray structure. The same is observed for the symmetric ligand in the 1TZ8 complex. The best docking solution reproduces all of the contacts present in the X-ray structure, but its rmsd (2.64 Å) is (only apparently) beyond the success cutoff. The native pose for the ligand in the 1YVF complex has a carboxylic group directly exposed to the solvent whose location is strongly penalized by the scoring function. Therefore, the native pose is not promoted to the top of the ranking list. Finally, the ligand in the 1SJ0 complex presents, in its central ring, a large substituent in the axial position and a small substituent in the equatorial position. This peculiar arrangement is not considered stable enough to be selected by CORINA as a representative conformation, and therefore, the docking algorithm fails in reproducing the native pose.

Another common error is the sulfonamide ligand in 1JD0, which was not docked correctly into human carbonic anhydrase XII using the rigid docking protocol (rmsd = 5.7 Å) but, rather surprisingly, was successfully docked (rmsd = 1.85 Å) when starting from scratch.

Ligand Conformational Sampling Errors. These errors are related to our ligand preparation step. In some cases the generated set of conformers does not properly cover the conformational space of a ligand. In the present study, the vast majority of the problematic cases are those in which the bound ligand in the experimental structure exhibits torsional angles whose values are away from those considered as ideal. These torsions are, to some extent, forced by the binding site environment. Since ideal angles are taken as rules in our conformer generator code ALFA, these particular conformations will not be present in the conformer population. Within the ADS, the ligands found in PDB entries 1Q1G, 1R58, 1UML, 1UNL, and 1UOU fall into this category.

To test whether or not this was, in fact, the source of error, these ligands were parametrized (GAFF, General Amber force field), immersed in cubic boxes of TIP3P water molecules, energy minimized (2000 steps of steepest descent followed by 2000 additional steps of Polak−Ribiere conjugate gradient), and simulated (constant temperature [300 K] and pressure [1 atm]) for 15 ns. Torsional angles were monitored over the whole trajectory and compared to those found in the crystal. In all five cases the relative population of conformers whose torsional angles were within 5° of the crystallographic structure was below 7%. This means that the native conformation is rarely sampled, and therefore, algorithms based on rules are very unlikely to generate near-native conformations.

Errors Due to Missing Water-Mediated Interactions. There are five cases of failure in which water molecules crucially mediate ligand−protein interactions (water-mediated bridges). As the usual protocol consists of eliminating, among other species, all water molecules present in a binding site, docking algorithms are unable to reproduce the native pose. This problem occurs in PDB entries 1GM8, 1G9V, 1GPK, 1HVY, and 1XM6 (Supporting Information, Figure S1) and was also found in the rigid ligand docking experiments. The ligand in complex 1GM8 has a β-lactam ring that interacts with residue Ser386 through a water-mediated bridge (WAT2460) and a benzylic amide buried at the bottom of the binding pocket. CRDOCK correctly reproduces the X-ray position of the latter moiety, while it places the β-lactam ring in an alternate location (rmsd = 3.37 Å). In the 1G9V complex, WAT916 mediates the interaction between Asn308 and the nitrogen atom of the ligand's amide group whereas WAT927 and WAT983 interact with the ligand's carboxylate. Both functional groups are wrongly positioned by CRDOCK in the absence of the water molecules (rmsd = 3.77 Å). The ligand in the 1GPK complex has a charged amino group interacting with a water molecule (WAT2529). When the water molecule is not considered as part of the binding site, the amino group is attracted by Glu199 and an alternative docking pose is favored (rmsd = 3.71 Å). The ligand in complex 1HVY has two different sources of error. First, the conformational sampling of the ligand fails because the closest conformer found is 1.47 Å away from the X-ray structure; second, the absence of two water molecules (WAT477 and WAT1017) precludes the correct orientation of the two carboxylic moieties at one end of the molecule (rmsd = 3.36). Finally, the ligand in the 1XM6 complex is unable to correctly orientate the oxazolidinone ring if WAT1009, which chelates the Zn^{2+} ion, is not present in the binding site. In its absence, the ligand's carbonyl group is strongly attracted by the Zn^{2+} ion (rmsd = 2.45).

The "missing water" problem could be easily solved in some cases using cGRILL to calculate the water affinity map within the binding site and place the relevant water molecule(s). These were correctly identified for both 1G9V and 1XM6, and upon incorporation into the protein structure, the rmsd of the ligands was notably reduced relative to the "dry" docking solution (from 3.77 to 1.51 Å and from 2.45 to 1.90 Å for 1G9V and 1XM6, respectively). For the remaining cases adding the water molecules helped only in part. In the 1HVY complex the ligand's carboxylate groups were correctly positioned upon addition of the water molecules, but the conformational problem still persisted (rmsd = 2.75 Å). Proper docking of the ligand in the 1GM8 complex improved significantly following incorporation of the water molecules; most of the contacts with the target were reproduced, but the rmsd was still 2.67 Å. On the other hand, for the same ligand in the "rigid" approach the rmsd decreased from 4.27 to 0.61 Å. Finally, for the 1GPK complex, addition of the relevant water molecules resulted in a pose with rmsd = 1.39 Å, although this solution was the second best on the ranking list.

Other Errors. Some compounds were classified as misdocked because of an rmsd slightly above 2.0 Å despite the fact that the top scoring pose faithfully reproduced the main interaction points that are observed within the binding site in the X-ray crystal structure. Five complexes belong to this category of "soft errors": 1HP0 (rmsd = 2.57 Å), 1N2V (rmsd = 2.86 Å), 1MMV (rmsd = 2.84 Å), 1XOQ (rmsd = 2.10 Å), and 2BM2 (rmsd = 2.25 Å). Finally, the ligands in complexes 1IG3, 1OQ5, 1P62, and 2BSM were not correctly positioned and no clear reason could be found to account for the deviations, which are presumably due to several combined factors such as insufficient or deficient ligand conformational sampling and/or inaccuracies in the scoring function.

3.2. Virtual Screening: Discriminating True Binders from Decoys. VS calculations were performed with the 40 targets and their associated sets of true binders and decoys from DUD. Global as well as family-wise statistics (AUCs from the ROC plots) were compiled and are summarized in Table 2. The average AUC values are comparable to those already

reported,[45] and there is a clear variability depending on the target. The 5 top scores are obtained for RXRα (AUC = 0.97), DHFR (AUC = 0.94), ER_{ago} (AUC = 0.86), trypsin (AUC = 0.84), and COX2 (AUC = 0.80). These targets are related neither functionally nor structurally, except for ER_{ago} and RXRα which belong to the family of nuclear hormone receptors (NHR). The same lack of obvious connection applies to the worst scoring targets, with the exception of the two NHR members PR and PPARγ: TK (AUC = 0.47), PR (AUC = 0.45), PPARγ (AUC = 0.42), ACE (AUC = 0.41), and AmpC (AUC = 0.32). The trend that CRDOCK performs better than average on folate enzymes and serine proteases has been observed with other docking programs.[45] In the case of kinases our method outperforms other docking tools reporting on the same data set[45] even though the average AUC value for this family is still below the global AUC average.

We took a closer view at the ROC curves (Supporting Information, Figure S2) to shed more light into those targets for which VS performance was worse than random: AChE (AUC = 0.49), TK (AUC = 0.47), PR (AUC = 0.47), ACE (AUC = 0.41), InhA (AUC = 0.48), and AmpC (AUC = 0.32). AmpC appears to be one of most problematic targets for all docking programs.[42] AChE[46] and TK[47] are two well-known enzymes in which the flexibility of some residues within the active site plays a crucial role in ligand recognition. As our docking tool does not include receptor flexibility, it is not entirely surprising that the VS protocol that we used fails in these cases, as do many other programs.[42] Although the global AUC values were below random for PR, ACE, and InhA, early enrichments (as detected in other studies)[42] were clearly apparent because 3/5, 4/9, and 13/17 true binders, respectively, were present on the top 0.5% of the rank-ordered list.

Different docking programs using DUD as a test set afforded values similar to those reported here for CRDOCK. Cross et al.[42] found average AUC values of 0.55 (DOCK), 0.59 (PhDock), 0.61 (FlexX), 0.63 (ICM), 0.66 (Surflex), and 0.72 (Glide). Finally, Marco et al.[48] reported a median AUC of 0.69 (very similar to CRDOCK) and Novikov et al.[44] an average AUC around 0.70.

3.3. GPCR: Pose Prediction and Virtual Screening. The original ADS was supplemented with five GPCR complexes, and self-docking experiments were carried out as described in the pose prediction section. The ligands in PDB entries 3PBL, 3UON, 2RH1, and 3RZE were correctly positioned within the binding site (rmsd $\leq$ 2.0 Å) in both the rigid and the flexible ligand docking tests. In the case of the μ opioid receptor, however, the first solution had rmsd = 2.44 Å, but it has to be borne in mind that in the crystal structure (PDB entry 4DKL) the funaltrexamine ligand is covalently bonded to Lys233. In the CRDOCK solution (Supporting Information, Figure S3) the main deviation lies precisely on the flexible chain that is covalently bonded to the amino group of Lys233 in the experimental structure, whereas the morphinane core reproduced the native pose accurately (rmsd = 1.27 Å). Therefore, although the overall rmsd is above the canonical cutoff value of 2.0 Å, the docking error can be considered "soft" in light of the conserved important interactions.

To explore the ability of CRDOCK to predict selectivity, we performed a simple cross-docking VS experiment using these five GPCR complexes, that is, every ligand was docked into every receptor to check whether the best score was assigned to the true ligand−receptor couple. In addition, by superimposing the seven transmembrane helices of all the GPCR studied we were able to calculate and compare the rmsd for all docking poses (Table 3). Without exception, the lowest rmsd was found for the native ligand in its own receptor. In addition, a very low rmsd at the M_2 receptor was found for doxepin, an H_1 receptor antagonist which is also known to bind with high affinity to muscarinic,[49] α_1-adrenergic,[50] and some mosquito dopamine[51] receptors. Regarding the ranking, the native ligands received the top scores at the β_2, M_2, and μ opioid receptors but those of the D_3 and H_1 receptors got the second best. At the D_3 receptor it was doxepin that appeared in the first position, and at H_1 it was not doxepin, as expected, but 3-quinuclidinyl benzilate, the prototypical anticholinergic agent with high selectivity for M_2 receptors. These findings raised a note of caution and prompted us to perform further studies.

Table 3. RMSD Values (in Angstroms, top) and Scores (kcal mol^{-1}, bottom) for Cross-Docking Studies on GPCRs

	ligands				
	β_2	D_3	H_1	M_2	μ
GPCR[a]					
β_2	**1.10**	3.81	5.67	4.58	5.42
D_3	3.80	**2.00**	5.10	6.80	7.18
H_1	4.29	3.98	**1.61**	4.48	8.11
M_2	4.67	4.60	**2.16**	**1.15**	5.93
μ	6.03	5.21	8.51	4.72	**1.27**
GPCR[b]					
β_2	**−101.3**	−86.1	−85.6	−82.8	−46.2
D_3	−83.9	**−94.0**	*−98.2*	−91.0	−73.7
H_1	−111.2	−97.3	**−117.5**	*−123.0*	−107.7
M_2	−87.8	−83.5	−104.7	**−115.4**	−57.2
μ	−61.6	−77.5	−72.7	−88.0	**−92.1**

[a]Numbers in bold along the diagonal highlight the lowest rmsd values for the ligand bound to its cognate receptor and, in addition, for the H_1 ligand (doxepin) bound to the M_2 receptor. [b]Numbers in bold along the diagonal highlight the binding energies for the cognate ligands, which are sometimes less favorable (in the same row) than for a noncognate ligand (values in italics).

When all true (as annotated in the DrugBank[36]) GPCR ligands studied were merged into a single set and used in an expanded VS experiment against the five GPCR, the AUC values from the obtained ROC curves were 0.50 (D_3), 0.53 (β_2), 0.56 (M_2), 0.71 (μ), and 0.61 (H_1). More informative, however, is the number of true binders that are recovered for each target at the top 10 of the ordered list: 6 are correctly identified for the β_2, D_3, and H_1 receptors, 5 for the M_2 receptor, and 2 for the μ opioid receptor.

Finally, when the set of ligands for each GPCR consisted of 30 decoys per true binder, the AUC from the resulting ROC curves were 0.50 (D_3), 0.56 (β_2), 0.64 (M_2), 0.80 (μ), and 0.82 (H_1). The average value of 0.67 compares very well with the findings reported above for the DUD. Nonetheless, it has to be noted that the number of true binders for each GPCR varied greatly, from 7 for the opioid receptor to 74 for the H_1 receptor (Supporting Information, Tables S1−S5), and this wide difference can be largely responsible for the diversity of the outcome. After completion of this work, we became aware of a recent compilation of 147 GPCR targets and a ligand library (agonists + antagonists) that included 39 decoy molecules for each true binder.[52] Application of the CRDOCK VS protocol to our selected GPCRs and this alternative compound

collection resulted in a small improvement for the β_2 receptor (0.1 AUC units), roughly the same performance for the D_3 receptor, and a slight decrease for μ (0.2 AUC units), H_1 (0.2 AUC units), and M_2 (0.1 AUC units) receptors. These differences, which amount to an average decrease in AUC from 0.67 to 0.56, can be expected due to the distinct way decoys and true ligands were selected and also to the fact that all receptors used are in the antagonist-bound conformation.

Altogether, these results are encouraging but indicate that to study selectivity among related GPCR with an acceptable degree of accuracy further improvements in CRDOCK and receptor preparation (e.g., incorporation of bound water molecules)[52] will be necessary.

3.3. Alternative Docking Protocols with Flexible Ligands. Two other alternatives were tested: our old in-house docking engine CDOCK and a different CRDOCK configuration.

From the results obtained for VS using DUD, CRDOCK represents a significant improvement over the rest (Table 1), which confirms the importance of the hybrid scoring function. However, performance on an individual target is in some cases strongly dependent on the selected scoring function. For example, in the case of the PDB code 1XGJ (β-lactamase AmpC) the AUC for the hybrid scoring function is as low as 0.32, but this figure is increased to 0.71when a force-field-based scoring function is used. Therefore and in agreement with other studies, we believe that trying to develop a universal scoring function[53,54] can be a daunting task indeed. Instead, more satisfactory results for the problem in hand can be achieved if a tailor-made target-dependent scoring function is used.

3.4. Benchmarking. The average time to perform the docking of a single flexible ligand using the reported combination of pieces that yielded the best performance is ~10 s on a 64-bit 3.3 GHz Intel Core i5 processor. We observed that the triangle matching approach could be used for around 91% of the compounds in the database. For the remaining ligands either exhaustive search or MCSA was used, and this took ~26 s on average to complete, leaving the docking average time in ~13 s. Therefore, with a modest cluster of 100 processors it should be possible to screen more than one-half million compounds per day in a typical VS campaign.

4. CONCLUSIONS

We introduce CRDOCK, a new computational tool that performs reasonably well in both pose prediction (docking) and true binder discrimination (VS). CRDOCK has demonstrated its abilities on two widely used benchmarking tests: the ADS for pose prediction and the DUD for VS. The docking failures found, some in common with other published reports, were analyzed in detail, and several solutions were found. In addition, five representative ligand−GPCR complexes were studied, and the results were in line with those obtained from DUD. Self-docking was always satisfactory with rmsd values below 2.0 Å, cross-docking was correct in 3 out of 5 cases, and the VS results provided encouraging results that support previous evidence[55] suggesting the feasibility of carrying out successful VS campaigns on pharmacologically important GPCR.[52,56] It is expected that future work on the remaining caveats will enhance CRDOCK performance.

Besides its accuracy, an additional advantage of CRDOCK is its reduced computational cost, once the conformational library for the ligands has been generated. It should be noted that this process is done only once, and the resulting conformers can be employed in different VS campaigns, so that millions of compounds can be screened thereafter using a relatively modest computational infrastructure. CRDOCK is open source and can be downloaded free of charge to noncommercial parties following registration at the CBM Bioinformatics Unit's web page (http://ub.cbm.uam.es/).

■ ASSOCIATED CONTENT

Supporting Information

Figures containing a structural depiction of docking errors, ROC curves corresponding to those targets for which VS performance was worse than random, and X-ray and docking solution for the ligand present in PDB entry 4DKL; tables listing the β_2, D_3, H_1, M_2, and μ binders used in the VS of GPCRs as well as the AUC values per DUD target. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: amorreale@cbm.uam.es.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by grants from CICYT (SAF2009-13914-C02-02 to F.G.) and Comunidad Autónoma de Madrid (S-BIO-0214-2006 [BIPEDD] and S2010-BMD-2457 [BIPEDD2] to A.M. and F.G.). A.M. and J.K. acknowledge financial support from Fundación Severo Ochoa through the AMAROUTO program and Ministerio de Economía y Competitividad (BFU2011-24595), respectively. R.G.-R. enjoyed a MICINN contract from "Programa de Personal Técnico y de Apoyo 2008", and A.C. is the recipient of FPU grant AP2009-0203 from the Ministerio de Educación. We are grateful to OpenEye Scientific Software, Inc. for providing us with an academic license for their software. The technical support and advice from the Bioinformatics team at CBMSO is gratefully acknowledged.

■ ABBREVIATIONS

AUC, area under the curve; BFGS, Broyden−Fletcher−Goldfarb−Shanno; DUD, Directory of Useful Decoys; MCSA, Monte Carlo simulated annealing; rmsd, root-mean-square deviation; ROC, receiver operating characteristic; VSDMIP, Virtual Screening Data Management on an Integrated Platform

■ REFERENCES

(1) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(2) Woltosz, W. S. If we designed airplanes like we design drugs. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 159−163.

(3) Kolb, P.; Irwin, J. J. Docking screens: right for the right reasons? *Curr. Top. Med. Chem.* **2009**, *9*, 755−770.

(4) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(5) Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins* **1990**, *8*, 195−202.

(6) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.
(7) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.
(8) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(9) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68*, 76−90.
(10) Pérez, C.; Ortiz, A. R. Evaluation of docking functions for protein-ligand docking. *J. Med. Chem.* **2001**, *44*, 3768−3785.
(11) Gil-Redondo, R.; Estrada, J.; Morreale, A.; Herranz, F.; Sancho, J.; Ortiz, A. R. VSDMIP: virtual screening data management on an integrated platform. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 171−184.
(12) Cabrera, Á. C.; Gil-Redondo, R.; Perona, A.; Gago, F.; Morreale, A. VSDMIP 1.5: an automated structure-and ligand-based virtual screening platform with a PyMOL graphical user interface. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 813−824.
(13) Cortes-Cabrera, A.; Gago, F.; Morreale, A. A reverse combination of structure-based and ligand-based strategies for virtual screening. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 319−327.
(14) Marshall, G. R. Limiting assumptions in structure-based design: binding entropy. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 3−8.
(15) Yuriev, E.; Agostino, M.; Ramsland, P. A. Challenges and advances in computational docking: 2009 in review. *J. Mol. Recognit.* **2011**, *24*, 149−164.
(16) Mysinger, M. M.; Shoichet, B. K. Rapid context-dependent ligand desolvation in molecular docking. *J. Chem. Inf. Model.* **2010**, *50*, 1561−1573.
(17) Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A. Target flexibility: an emerging consideration in drug discovery and design. *J. Med. Chem.* **2008**, *51*, 6237−6255.
(18) Kokh, D. B.; Wade, R. C.; Wenzel, W. Receptor flexibility in small-molecule docking calculations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*; Wiley: New York, 2011; Vol. *1*, pp 298−314.
(19) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−741.
(20) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(21) Liebeschuetz, J.; Cole, J.; Korb, O. Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test. *J. Comput.-Aided Mol. Des.* **2012**, Epub ahead of print.
(22) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 1−14.
(23) *Corina Molecular Networks*; GmbH Computerchemie Langemarckplatz 1, E., Germany, 2000.
(24) Stewart, J. J. MOPAC: a semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.
(25) Case, D. A.; Darden, T. A.; Cheatham, I., T.E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlam, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER 8*; University of San Francisco: San Francisco, 2004.
(26) Gil-Redondo, R. Master Thesis. UNED, Madrid, 2006.
(27) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.
(28) Gschwend, D. A.; Kuntz, I. D. Orientational sampling and rigid-body minimization in molecular docking revisited: on-the-fly optimization and degeneracy removal. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 123−132.
(29) Wang, J.; Kollman, P. A.; Kuntz, I. D. Flexible ligand docking: a multistep strategy approach. *Proteins* **1999**, *36*, 1−19.
(30) Nelder, J. A.; Mead, R. A simplex method for function minimization. *Comput. J.* **1965**, *7*, 308.
(31) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Daniel, T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(32) Fan, H.; Schneidman-Duhovny, D.; Irwin, J. J.; Dong, G.; Shoichet, B. K.; Sali, A. Statistical Potential for Modeling and Ranking of Protein-Ligand Interactions. *J. Chem. Inf. Model.* **2011**, *51*, 3078−3092.
(33) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534−2547.
(34) Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **2007**, *35*, W522−W525.
(35) DeLano, W. L. *The PyMOL molecular graphics system*; Schrodinger Inc.: Cambridge, MA, 2002.
(36) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035.
(37) Irwin, J. J.; Shoichet, B. K. ZINC-a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(38) Jorgensen, W. L. *QikProp*; Schrödinger LLC: New York, 2006.
(39) Ruppert, J.; Welch, W.; Jain, A. N. Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci.* **1997**, *6*, 524−533.
(40) Baber, J. C.; Thompson, D. C.; Cross, J. B.; Humblet, C. GARD: a generally applicable replacement for RMSD. *J. Chem. Inf. Model.* **2009**, *49*, 1889−1900.
(41) Li, X.; Li, Y.; Cheng, T.; Liu, Z.; Wang, R. Evaluation of the performance of four molecular docking programs on a diverse set of protein-ligand complexes. *J. Comput. Chem.* **2010**, *31*, 2109−2125.
(42) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455−1474.
(43) Spitzer, R.; Jain, A. N. Surflex-Dock: Docking benchmarks and real-world application. *J. Comput.-Aided Mol. Des.* **2012**, Epub ahead of print.
(44) Novikov, F. N.; Stroylov, V. S.; Zeifman, A. A.; Stroganov, O. V.; Kulkov, V.; Chilov, G. G. Lead Finder docking and virtual screening evaluation with Astex and DUD test sets. *J. Comput.-Aided Mol. Des.* **2012**, Epub ahead of print.
(45) von Korff, M.; Freyss, J.; Sander, T. Comparison of ligand- and structure-based virtual screening on the DUD data set. *J. Chem. Inf. Model.* **2009**, *49*, 209−231.
(46) Alberts, I. L.; Todorov, N. P.; Dean, P. M. Receptor flexibility in de novo ligand design and docking. *J. Med. Chem.* **2005**, *48*, 6585−6596.
(47) Fischer, B.; Merlitz, H.; Wenzel, W. Increasing diversity in in-silico screening with target flexibility. *Comput. Life Sci.* **2005**, 186−197.
(48) Neves, M. A.; Totrov, M.; Abagyan, R. Docking and scoring with ICM: the benchmarking results and strategies for improvement. *J. Comput.-Aided Mol. Des.* **2012**, Epub ahead of print.

(49) Ehlert, F. J.; Delen, F. M.; Yun, S. H.; Liem, H. A. The interaction of amitriptyline, doxepin, imipramine and their N-methyl quaternary ammonium derivatives with subtypes of muscarinic receptors in brain and heart. *J. Pharmacol. Exp. Ther.* **1990**, *253*, 13−19.
(50) Richelson, E.; Nelson, A. Antagonism by antidepressants of neurotransmitter receptors of normal human brain in vitro. *J. Pharmacol. Exp. Ther.* **1984**, *230*, 94−102.
(51) Meyer, J. M.; Ejendal, K. F.; Avramova, L. V.; Garland-Kuntz, E. E.; Giraldo-Calderon, G. I.; Brust, T. F.; Watts, V. J.; Hill, C. A. A "Genome-to-lead" approach for insecticide discovery: pharmacological characterization and screening of *Aedes aegypti* D(1)-like dopamine receptors. *PLoS Negl. Trop. Dis.* **2012**, *6*, e1478.
(52) Gatica, E. A.; Cavasotto, C. N. Ligand and decoy sets for docking to G protein-coupled receptors. *J. Chem. Inf. Model.* **2012**, *52*, 1−6.
(53) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **2008**, *153* (Suppl 1), S7−26.
(54) Tarasov, D.; Tovbin, D. How sophisticated should a scoring function be to ensure successful docking, scoring and virtual screening? *J. Mol. Model.* **2009**, *15*, 329−341.
(55) Carlsson, J.; Coleman, R. G.; Setola, V.; Irwin, J. J.; Fan, H.; Schlessinger, A.; Sali, A.; Roth, B. L.; Shoichet, B. K. Ligand discovery from a dopamine D3 receptor homology model and crystal structure. *Nat. Chem. Biol.* **2011**, *7*, 769−778.
(56) Mysinger, M. M.; Weiss, D. R.; Ziarek, J. J.; Gravel, S.; Doak, A. K.; Karpiak, J.; Heveker, N.; Shoichet, B. K.; Volkman, B. F. Structure-based ligand discovery for the protein-protein interface of chemokine receptor CXCR4. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 5517−5522.

Supporting Information

CRDOCK: an ultrafast multipurpose protein-ligand docking tool

Álvaro Cortés Cabrera[1,2], Javier Klett[2], Helena Gomes[2], Almudena Perona[2,3], Rubén Gil-Redondo[2,3], Sandrea M. Francis[4], Eva M. Priego[5], Federico Gago[1], and Antonio Morreale[2]

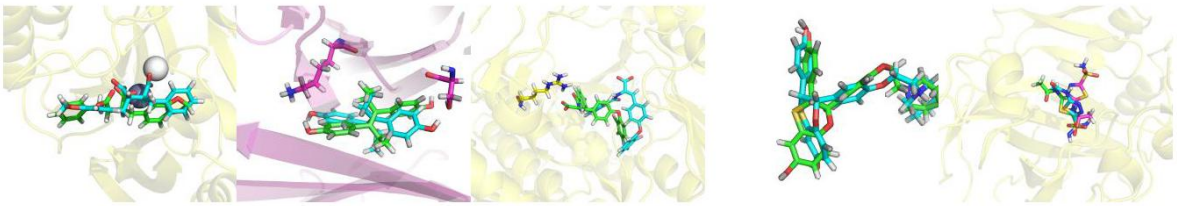
[1]Departamento de Farmacología, Universidad de Alcalá, E-28871 Alcalá de Henares, Madrid, Spain
[2]Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Campus UAM, c/Nicolás Cabrera 1, E-28049 Madrid, Spain
[3]SmartLigs Bioinformática S.L., Fundación Parque Científico de Madrid, c/ Faraday, 7. Campus de Cantoblanco UAM, E-28049 Madrid, Spain
[4]Instituto de Biomedicina de Valencia (IBV-CSIC), c/ Jaime Roig 11, E-46010 Valencia, Spain
[5]Instituto de Química Médica (CSIC), c/ Juan de la Cierva 3, E-28006 Madrid, Spain

Corresponding author: Antonio Morreale. Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Campus UAM, c/ Nicolás Cabrera 1, E-28049 Madrid, Spain. Email: amorreale@cbm.uam.es. Telephone number: + 34 91196 4633. Fax number: + 34 91196 4422.

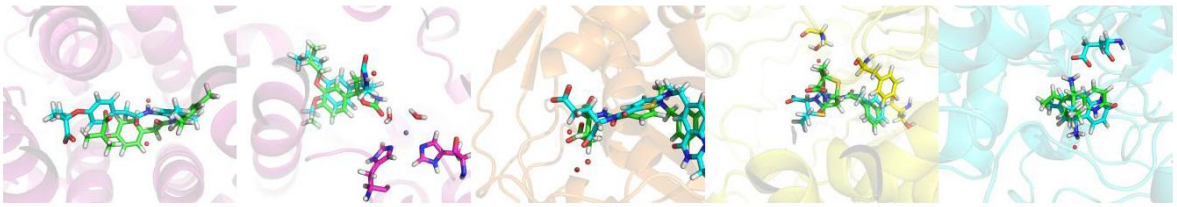
Figure S1.Structural depiction of docking errors. In all the pictures, the ligand X-ray structure is represented with the carbon atoms in blue, while the docking solution is shownwith the carbon atoms in green. Water molecules are shown as small red spheres.

Common errors. 1JJE (zinc ion in white) and 1TZ8: the docking solution is rotated 180° compared to the X-ray structure; 1YVF: the X-ray structure exposes a carboxylic group to the solvent, a solution that is highly penalized by the scoring function; 1SJ0: different arrangement of substituents at the central ring of the molecule; 1JD0: not properly docked in rigid ligand docking while successfully in flexible ligand docking.

Errors due to missing water-mediated interactions. 1GM8 and 1XM6: the docking solution was thereafter correctly located within the binding site upon inclusion of the water molecules; 1HVY, 1GM8 and 1GPK: subsequent addition of water molecules helped only in part to recover the X-ray ligand pose (see main text).



1JJE 1TZ8 1YVF 1SJ0 1JD0



1G9V 1XM6 1HVY 1GM8 1GPK

Figure S2. ROC curves corresponding to those targets for which VS performance was worse than random. AChE, TK, and AmpC targets are well-known difficult examples for rigid protein docking. For PR, ACE and InhA, early enrichments (red circles) were achieved.

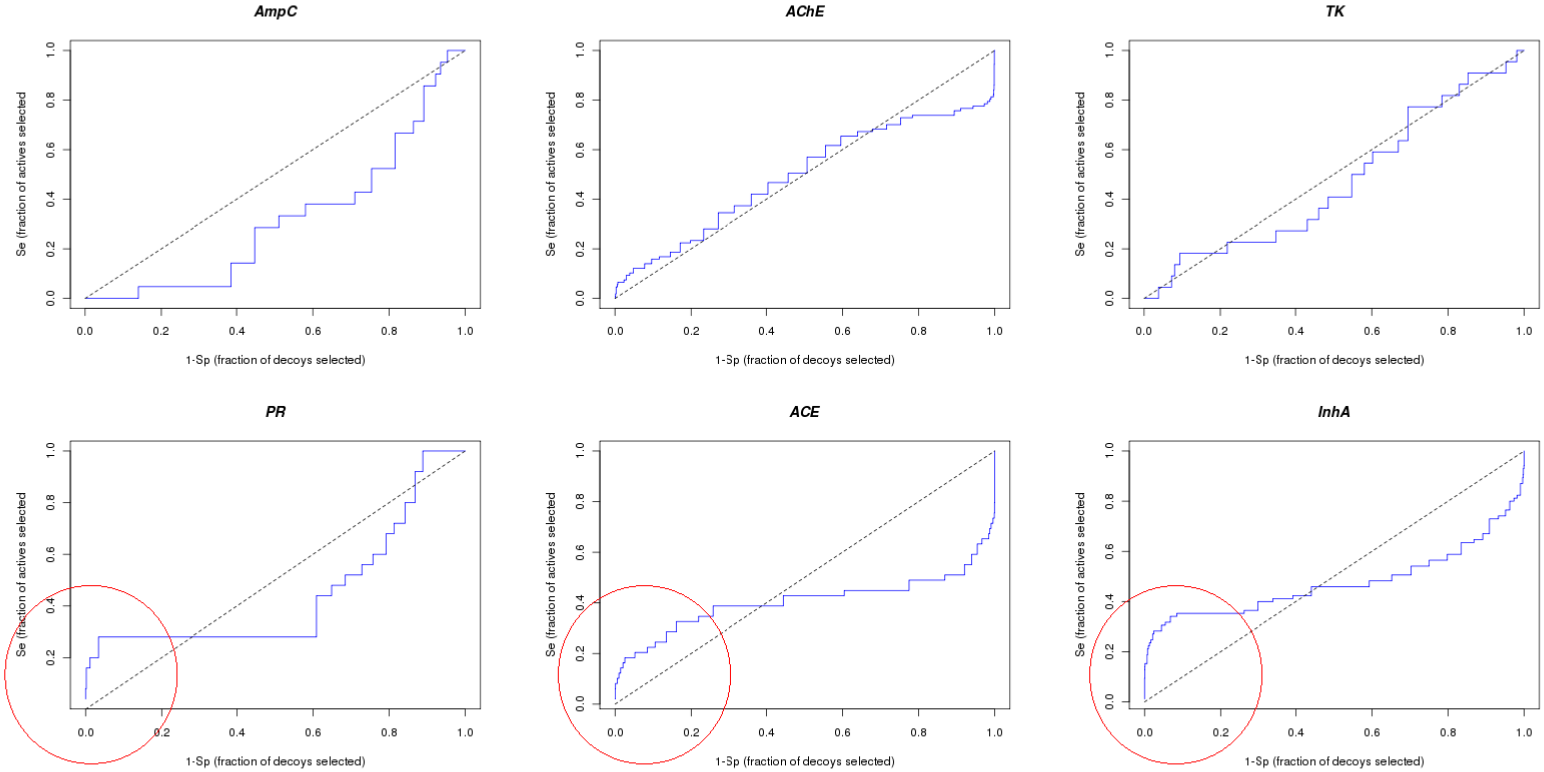


Figure S3. X-ray ligand pose (carbon atoms in grey) and docking solution (carbon atoms in green) corresponding to PDB entry 4DKL (μ opioid receptor, cyan): the main interactions are reproduced but the flexible chain deviates from the X-ray pose (covalently bonded to Lys233).

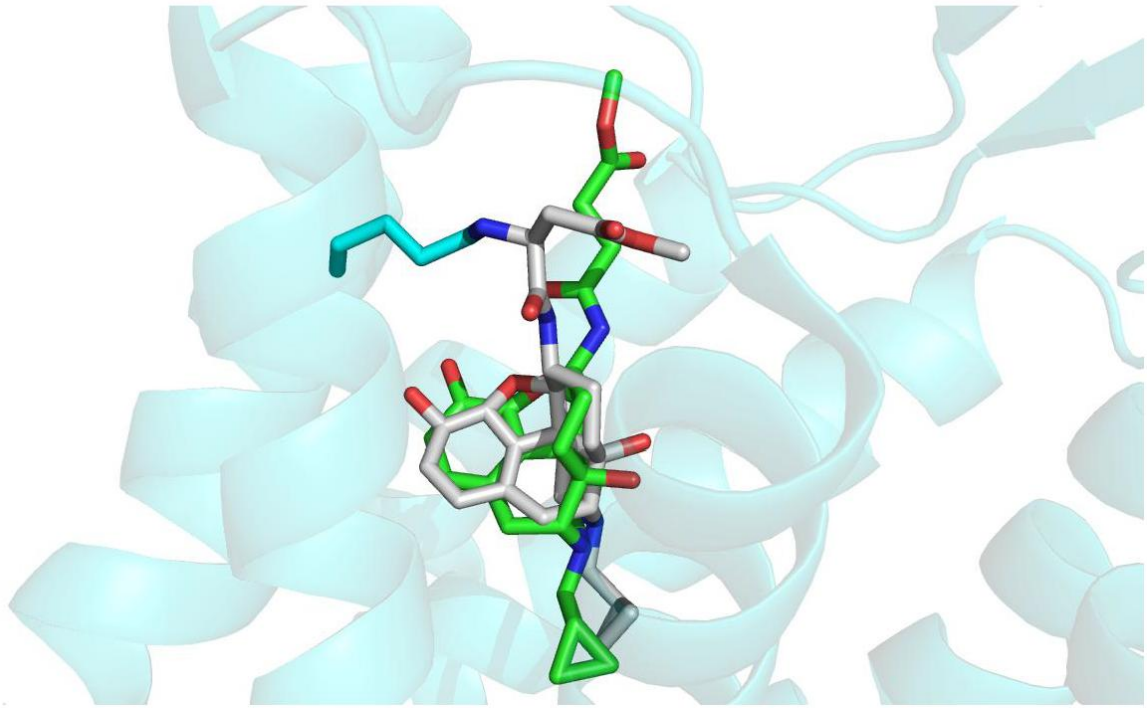


Table S1. β_2 binders from DrugBank used in the VS experiments

(2S)-1-(9*H*-Carbazol-4-yloxy)-3-(isopropylamino)propan-2-ol
Acebutolol
Alprenolol
Arbutamine
Arformoterol
Bambuterol
Betaxolol
Bevantolol
Bisoprolol
Bitolterol
Bopindolol
Bupranolol
Carteolol
Carvedilol
Clenbuterol
Desipramine
Dipivefrin
Dobutamine
Droxidopa
Ephedra
Epinephrine
Fenoterol
Formoterol
Isoproterenol
Labetalol
Levobunolol
Metipranolol
Metoprolol
Nadolol
Nebivolol
Norepinephrine
Orciprenaline
Oxprenolol
Penbutolol
Phenoxybenzamine
Pindolol
Pirbuterol
Procaterol
Propranolol
Pseudoephedrine
Ritodrine
Salbutamol
Salmeterol
Sotalol
Terbutaline
Timolol

Table S2. D_3 binders from DrugBank used in the VS experiments

Amisulpride
Apomorphine
Aripiprazole
Bromocriptine
Cabergoline
Chlorprothixene
Clozapine
Domperidone
Dopamine
Haloperidol
Levodopa
Lisuride
Methotrimeprazine
Olanzapine
Paliperidone
Pergolide
Pimozide
Pramipexole
Quetiapine
Remoxipride
Risperidone
Ropinirole
Rotigotine
Sulpiride
Yohimbine
Ziprasidone

Table S3. H_1 binders from DrugBank used in the VS experiments

Aceprometazine	Fexofenadine
Alcaftadine	Flunarizine
Amitriptyline	Histamine phosphate
Antazoline	Hydroxyzine
Aripiprazole	Imipramine
Astemizole	Isothipendyl
Azatadine	Ketotifen
Azelastine	Levocabastine
Benzquinamide	Loratadine
Benztropine	Maprotiline
Bepotastine	Meclizine
Betahistine	Mepyramine
Bromodiphenhydramine	Mequitazine
Brompheniramine	Methdilazine
Buclizine	Methotrimeprazine
Carbinoxamine	Mianserin
Cetirizine	Mirtazapine
Chlophedianol	Nortriptyline
Chloropyramine	Olanzapine
Chlorpheniramine	Olopatadine
Chlorpromazine	Orphenadrine
Chlorprothixene	Paliperidone
Cinnarizine	Pemirolast
Clemastine	Phenindamine
Clozapine	Pheniramine
Cyclizine	Promazine
Cyproheptadine	Promethazine
Desipramine	Propiomazine
Desloratadine	Quetiapine
Dexbrompheniramine	Risperidone
Dimenhydrinate	Terfenadine
Dimethindene	Tolazoline
Diphenhydramine	Trazodone
Diphenylpyraline	Trimeprazine
Doxepin	Trimipramine
Doxylamine	Tripelennamine
Emedastine	Triprolidine
Epinastine	Ziprasidone
Escitalopram	

Table S4. M_2 binders from DrugBank used in the VS experiments

Amitriptyline	Methotrimeprazine
Anisotropine methylbromide	Methylscopolamine
Aripiprazole	Metixene
Atropine	Metocurine
Benzquinamide	Mivacurium
Bethanechol	Nicardipine
Brompheniramine	Nortriptyline
Carbachol	Olanzapine
Chlorprothixene	Oxybutynin
Clozapine	Oxyphencyclimine
Cocaine	Pancuronium
Cryptenamine	Paroxetine
Cyproheptadine	Pilocarpine
Darifenacin	Pipecuronium
Desipramine	Procyclidine
Dicyclomine	Promazine
Dimethindene	Promethazine
Diphenidol	Propiomazine
Disopyramide	Quetiapine
Doxacurium chloride	Rocuronium
Doxepin	Scopolamine
Ethopropazine	Solifenacin
Fesoterodine	Succinylcholine
Flavoxate	Tiotropium
Gallamine triethiodide	Tolterodine
Homatropine methylbromide	Tridihexethyl
Hyoscyamine	Trifluprom azine
Imipramine	Trihexyphenidyl
Ipratropium bromide	Tropicamide
Maprotiline	Ziprasidone

Table S5.μ binders from DrugBank used in the VS experiments

Alvimopan
Butorphanol
Diprenorphine
Nalbuphine
Naloxone
Naltrexone
Pentazocine

Table S6. AUC results for individual DUD targets.

Target	AUC[a]
MR	0.79
ACE	0.41
PNP	0.61
HIVPR	0.79
TK	0.47
Thrombin	0.76
FXa	0.74
AMPC	0.32
HMGR	0.64
COMT	0.63
HSP90	0.74
P38	0.77
COX1	0.55
ACHE	0.49
GPB	0.73
COX2	0.80
ALR2	0.51
ER_AGO	0.86
PR	0.47
AR	0.78
ADA	0.76
GART	0.72
DHFR	0.89
CDK2	0.59
EGFR	0.72
RXR	0.97
PDE5	0.63
SAHH	0.79
PARP	0.57
PPAR	0.61
NA	0.53
INHA	0.48
GR	0.71
HIVRT	0.67
SRC	0.67
TRYPSIN	0.84
VEGFR2	0.75
FGFR1	0.56
PDGFRB	0.51
ER_ANTAGO	0.70

[a]Area under the curve of the generated ROC plots.

Article III

A reverse combination of structure-based and ligand-based strategies for virtual screening

Background and author's contribution

The problem of diversity in chemical libraries intended for VS has been present since the ancient times of CADD back in the 80s. This is so because these compound collections tend to be built with purchasability or chemical accessibility in mind and, very often, they are obtained from external providers who ensure that enough amount of product can be acquired to perform an activity confirmation test. Tailor-made libraries can be defined as well if enough information on the target is available, *e.g.* for metalloenzymes, several chemical moieties such as carboxylic acids or thiol groups have a higher chance to bind and therefore, an exclusive library of compounds possessing these characteristics may be constructed.

Anyway, whatever the source of the molecules is, the amount of time needed to perform the VS is directly proportional to the number of molecules in the collection, and hence, a rational design of the library to avoid redundancy and problematic/toxic compounds is desirable.

The next manuscript describes a simple methodology to compress the chemical space of any virtual library minimizing the loss of diversity through the use of smaller chemical entities. The chemical space shrinks significantly when only small fragments are considered and therefore these fragments may be of use to ensure diversity at an affordable computational cost. In addition, to obtain a complete coverage of the *drugs* chemical space, *i.e.* more complex molecules with more *drug-like* properties, we added an additional *expansion* step, which selects all molecules in the database that map directly to the fragments in that part of chemical space.

The main author of the manuscript designed the protocol, wrote the first draft of the manuscript and implemented the necessary scripts and applications for the analyses except for the original docking program (CDOCK) and VSDMIP platform, since this methodology was developed prior to the new docking program.

J Comput Aided Mol Des (2012) 26:319–327
DOI 10.1007/s10822-012-9558-x

A reverse combination of structure-based and ligand-based strategies for virtual screening

Álvaro Cortés-Cabrera · Federico Gago · Antonio Morreale

Received: 18 November 2011 / Accepted: 24 February 2012 / Published online: 7 March 2012

Abstract A new approach is presented that combines structure- and ligand-based virtual screening in a reverse way. Opposite to the majority of the methods, a docking protocol is first employed to prioritize small ligands ("fragments") that are subsequently used as queries to search for similar larger ligands in a database. For a given chemical library, a three-step strategy is followed consisting of (1) *contraction* into a representative, non-redundant, set of fragments, (2) *selection* of the three best-scoring fragments docking into a given macromolecular target site, and (3) *expansion* of the fragments' structures back into ligands by using them as queries to search the library by means of fingerprint descriptions and similarity criteria. We tested the performance of this approach on a collection of fragments and ligands found in the ZINC database and the directory of useful decoys, and compared the results with those obtained using a standard docking protocol. The new method provided better overall results and was several times faster. We also studied the chemical diversity that both methods cover using an in-house compound library and concluded that the novel approach performs similarly but at a much smaller computational cost.

Keywords Fragment screening · Structure-based virtual screening · Ligand-based virtual screening · Docking · Drug design

Electronic supplementary material The online version of this article (doi:10.1007/s10822-012-9558-x) contains supplementary material, which is available to authorized users.

Á. Cortés-Cabrera · F. Gago
Departamento de Farmacología, Universidad de Alcalá, 28871 Alcalá de Henares, Madrid, Spain

Á. Cortés-Cabrera · A. Morreale (✉)
Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC/UAM), Campus UAM, c/Nicolás Cabrera 1, 28049 Madrid, Spain
e-mail: amorreale@cbm.uam.es

Introduction

The publication of Abbott's seminal paper describing the SAR (structure–activity relationships) by nuclear magnetic resonance (NMR) method [1] introduced the *fragment* as a new concept within the field of drug discovery, shifting the emphasis from the paradigmatic more conventional ligands (in terms of size and affinity) to smaller pieces. These "fragments" are usually endowed with reduced affinities but are better suited for chemical modifications aimed at producing novel drug candidates. Since its onset, the technique has experienced a soaring success in large pharma, small biotechs, and academia [2]. Fragments can sample chemical space more effectively than regular ligands do [3] and fragment docking clearly outperforms traditional high-throughput screening in terms of hit rates [4, 5]. Parallel to the purposeful deployment of customized software, some other computational techniques have been adapted to handle fragments, in particular those that attempt to yield ligands by starting off from these building blocks. Fragments can evolve virtually (by adding different chemical decorations), be linked (to join two or more fragments that occupy different regions of the binding site), self-assemble (through direct bond formation between different reacting fragments), and/or be optimized to better fulfill drug-like properties. Fragments are usually docked and scored, but due to the fact that their volumes are smaller than the binding site cavity erroneous binding modes can be obtained. Besides, scoring functions need to be fine-tuned as they are parameterized for much larger

molecular entities. Nonetheless, despite these deficiencies, much progress has been made in the field and fragment-based ligand design (FBLD) has become a routine tool nowadays. Fragments can be designed de novo or obtained from ligand databases by chemical dissociation (i.e. fragmentation).

Virtual screening (VS) techniques that rely on the structure of either the macromolecular receptor (SBVS) or a known ligand (LBVS) use chemical libraries to search for hits that can then be transformed into leads. But due to the huge amount of ligands that can be found in today's databases (e.g. 13 million in ZINC [6]), it is impractical, as well as inefficient, to perform lengthy full docking studies. To speed up the process, it is customary to employ a series of computational cost-effective filters to narrow down the number of molecules that will be subjected to the demanding tasks of docking and scoring. Lipinki's rule of five [7] and/or other physico-chemical-based principles can be used as filters, as well as pharmacophoric hypotheses. LBVS can also be employed by taking some known ligands as templates although the outcome may be devoid of novelty because the resulting molecules tend to resemble the original queries.

In order to overcome this drawback, we designed a new protocol that combines SBVS and LBVS in a reverse order, that is, fragment SBVS serves as a previous filter to LBVS. This greatly reduces the computing time while maintaining the chemical diversity that is contained in the original library. For a given compound collection the new three-step strategy performs the following tasks: (1) *contracts* the database into a representative, non-redundant, set of fragments that are used for docking against the target of interest, (2) *selects* the three best-scoring fragments; and (3) *expands* the structure of the fragments back into ligands by employing the fragments as queries to search the database using a fingerprints description and a similarity criterion.

To test this new approach we first demonstrated that our docking tool is able to reproduce the experimental poses for a set of receptor-bound fragments in complexes of known 3D structure. Next, we applied the three-step procedure to the "fragment-like subset" in the ZINC database [6] using the 40 macromolecular targets contained in the directory of useful decoys (DUD) [8] and our CGRID/CDOCK docking tool [9]. The performance of the model was assessed by means of the area under the curve (AUC) of the generated receiver operating characteristics (ROC) plots. Finally, a comparison was made between the chemical space covered by the hits obtained from a standard VS protocol based on small-molecule docking and that provided by the top-ranking docked fragments generated from these same molecules.

Methods

Fragment definition

A fragment is defined as a molecular entity endowed with the following properties: (a) log $P \leq 2.5$, (b) molecular weight $\leq$250 Da, and (c) <6 rotatable bonds.

Fragment docking

Thirty-four complexes [10] from the ASTEX diverse set [11] in which the ligands fulfill the above fragment definition criteria were used as a test set. The fragments were extracted from the complexes and converted to SMILES [12] strings using OpenBABEL [13]. Then, our standard docking workflow was followed:

1. For the ligands: (a) conversion from SMILES to 3D MOL2 using CORINA [14], (b) atomic charge calculations with MOPAC [15] (AM1 ESP method) on every single structure provided by CORINA; and c) atom type assignment according to the AMBER force field [16] and conformational analysis using ALFA [17]. The protonation and tautomeric states for some of the ligands were manually adjusted (Fig. S1, Supplementary Information).
2. The receptors, including those water molecules and metal ions essential for ligand binding (Table S1, Supplementary Information), were prepared using pdb2pqr [18] and adapted to the AMBER force field by using 250 steps of steepest descent followed by 2,000 steps of Polak-Ribiere conjugate gradient energy minimization.
3. Docking of ligands and fragments was performed with our in-house CGRID/CDOCK tool: (a) definition of the binding site as the space delimited by the axis-parallel box containing the co-crystallized ligand, augmented by 5 Å in each axis direction, (b) CGRID calculation of protein interaction fields (a 12–6 Lennard–Jones term and an electrostatic term modeled with a sigmoidal dielectric screening function) covering the binding site (0.5 Å spacing in all directions) using common atom probes (C, N, O, S, P, H, F, Cl, Br, and I), (d) exhaustive exploration by CDOCK of the location and orientation of each fragment within the binding site by positioning their centers of mass on grid points and performing discrete rotations of 27° on each axis, (e) energy evaluation of each pose by the molecular mechanics force-field scoring function, as implemented in CDOCK, and (f) selection of the best-scoring pose for each fragment as the docking solution.

The standard criterion to validate the ability to predict the native pose was the root-mean-square deviation (rmsd) of the heavy atoms between the docking solution and the

native conformation for each fragment. In molecular docking, those poses within 2 Å from the experimental structure are usually considered as correct solutions. In the case of fragments, due to their reduced size, this cutoff is customarily decreased to 1.5 Å. Each docking experiment was run 20 times, and the reported rmsd corresponds to the average value. The success rate was defined as the percentage of poses having an average rmsd below 1.5 Å.

Fragment screening

The fragments used here were those belonging to the "fragment-like subset" in the ZINC database. Due to different conversion problems, not all the SMILES codes were able to produce the corresponding 3D structures, and for this reason only 6,183 were employed out of the original 7,106 strings. As receptors, we used the 40 structures comprising the DUD dataset (Table 1) and these were prepared for docking as explained before for the ASTEX test set.

Initially, all the fragments were docked (SBVS) using CGRID/CDOCK on each DUD target and only the best-scoring pose for each one was retained. Then, the three higher-ranking fragments for each target were selected and converted to MACCS fingerprints [19]. The LBVS protocol consisted of using these MACCS as queries to screen the DUD sets of real binders and decoys using the Tanimoto coefficient (Tc) index as the score. Performance is reported as the AUCs corresponding to the best experiment out of the three performed for each target (one for each of the three best fragments/target). The ligand list retrieved from each of these three fragments, at least in theory, should belong to different regions of chemical space. Thus, repeating the experiments 3 times is aimed at increasing the chemical diversity of the results. Test calculations with more than three fragments did not result in better coverage. In a real-world situation the top-scoring compounds for each of the three lists should be selected for testing, so as to improve the likelihood of finding new hits.

Comparative test

Fifty-two thousand two hundred and thirty-one molecules from an in-house chemical library that were ranked as possible hits in a standard SBVS campaign were decomposed into fragments with our tailor-made program, based on the chemistry development kit (CDK) [20], that makes use of the *exhaustive fragmentator* tool to break all rotatable bonds and generate fragments with at least 5 heavy atoms. A total of 1,137,482 fragments were thus extracted and then clustered using the stochastic clustering algorithm that is implemented in the SUBSET program [21]. MACCS fingerprints represented the fragments and a maximum Tc similarity index of 0.6 was used as a cutoff. This procedure, which is the same that was used in the ZINC database to obtain the so-called "fragment-like subset", yielded 2,540 non-redundant fragments. The new protocol then proceeded as follows:

1. Fragment docking and scoring with CGRID/CDOCK;
2. Selection of the best-scoring pose for each fragment and of the three best-scoring fragments;
3. Comparison between the chemical spaces covered by the fragments and by the parent compounds. The overlap (O, Eq. 1) between both spaces was obtained by:
   - an all-versus-all comparison between the 2,540 non-redundant fragments and the 52,231 parent compounds, and calculation of the Tcs among them as a control test;
   - selecting the top 1% compounds from both protocols and calculating the Tcs among them; and finally,
   - selecting the top 1% compounds from the standard protocol and the three best-scoring fragments, and calculating the Tcs among them.

$$O = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{Tc}\ (i,j) \tag{1}$$

where O, the overlap, is obtained as the averaged sum of the Tcs; m is the total number of ligands (520 [1% of 52,231]) and n the total number of fragments (either 24 [1% of 2,540] or only 3), and i and j are indices. To perform these comparisons ligands and fragments were represented using chemically advanced template search (CATS) descriptors [22].

Results and discussion

Fragment docking versus ligand docking

We first performed a "self-docking experiment" to test the accuracy of our CGRID/CDOCK docking engine when working with fragments. On average, for regular ligands (Fig. 1a), CGRID/CDOCK was able to reproduce the pose found in the X-ray crystal structure of the complex within an rmsd of 2.0 Å with a success rate of 75–80% (unpublished results). For fragments, and considering 1.5 Å as the cutoff value, we obtained a success rate of 80% with an average rmsd value of 0.88, in good consonance with recent studies [23, 24].

Fragment screening using ZINC fragments and DUD targets

Next, we tested the *contraction/selection/expansion* approach (Fig. 1b):

Table 1 AUCs values corresponding to each individual target depending on the method used: FBP (the fragment-based protocol presented here) and SP (standard ligand docking protocol)

Target[a]	FBP[b]	SP[c]	Target[a]	FBP[b]	SP[c]
ACE	0.79	0.63	HIVRT	0.53	0.61
AChE	0.45	0.73	HMGR	0.11	0.57
ADA	0.60	0.65	HSP90	0.83	0.69
ALR2	0.57	0.57	INHA	0.57	0.40
AMPC	0.36	0.58	MR	0.82	0.78
AR	0.76	0.65	NA	0.77	0.75
CDK2	0.68	0.52	P38	0.75	0.50
COMT	0.71	0.87	PARP	0.69	0.65
COX1	0.24	0.53	PDE5	0.67	0.78
COX2	0.51	0.67	PDGFRB	0.58	0.24
DHFR	0.67	0.49	PNP	0.53	0.60
EGFR	0.77	0.52	PPARγ	0.90	0.43
ER_{ago}	0.97	0.64	PR	0.76	0.49
ER_{antago}	0.75	0.81	RXRα	0.96	0.92
FGFR1	0.63	0.31	SAHH	0.51	0.82
FXa	0.61	0.55	SRC	0.53	0.48
GART	0.70	0.55	Thr	0.60	0.67
GPB	0.69	0.83	TK	0.68	0.59
GR	0.78	0.61	Trypsin	0.63	0.66
HIVPR	0.28	0.33	VEGFR2	0.71	0.41
					Averages
				SP	0.60
				FB[d]	0.54
				FB[e]	0.64

The average values appear in the last three rows

[a] *ACE* angiotensin-converting enzyme, *AChE* acetylcholinesterase, *ADA* adenosine deaminase, *ALR2* aldose reductase, *AmpC*, AmpC β-lactamase, *AR* androgen receptor, *CDK2* cyclin-dependent kinase 2, *COMT* catechol *O*-methyltransferase, *COX-1* cyclooxygenase-1, *COX-2*, cyclooxygenase-2, *DHFR* dihydrofolate reductase, *EGFR* epidermal growth factor receptor, ER_{ago} estrogen receptor (agonist-bound conformation), ER_{antago} estrogen receptor (antagonist-bound conformation), *FGFR1* fibroblast growth factor receptor kinase, *FXa* factor Xa, *GART* glycinamide ribonucleotide transformylase, *GP*β glycogen phosphorylase β, *GR* glucocorticoid receptor, *HIVPR* HIV protease, *HIVRT* HIV reverse transcriptase, *HMGR* hydroxymethylglutaryl-CoA reductase, *HSP90* human heat shock protein 90, *INHA* enoyl ACP reductase, *MR* mineralocorticoid receptor, *NA* neuraminidase, *P38 MAP* P38 mitogen activated protein, *PARP* poly(ADP-ribose) polymerase, *PDE5* phosphodiesterase 5, *PDGFRB* platelet derived growth factor receptor kinase, *PNP* purine nucleoside phosphorylase, *PPAR*γ peroxisome proliferator activated receptor γ, *PR* progesterone receptor, *RXR*α retinoic X receptor α, *SAHH* S-adenosyl-homocysteine hydrolase, *SRC* tyrosine kinase SRC, *Thr* thrombin, *TK* thymidine kinase, *VEGFR2* vascular endothelial growth factor receptor

[b] Fragment-based protocol presented in this paper

[c] Standard protocol

[d] The average was calculated with the original collection of fragments (not properly representing NHR binders' fragments)

[e] The average was calculated with the appropriate fragments to represent NHR binders (see Fig. 4)

(a) *Contract* the ligand library into a representative, non-redundant set of fragments that are used for docking;
(b) *select* the three best-scoring fragments; and
(c) *expand* these fragments by searching the database for ligands containing similar substructures.

For comparison purposes, SBVS was also carried out with the original ligands (true binders + decoys). Despite the fact that the total number of ligands was less than the number of fragments, the time and computational resources required for this procedure were generally larger than in the new fragment-based approach (see below). Overall, the average AUC for the two protocols was very similar ($\sim$0.6, Table 1) because, although the new method was superior for 62% of the targets (26 out of 42), it proved inferior in a few others. To analyze in more detail these differences in performance the targets were grouped into families (Fig. 2).

This allowed us to see that the new methodology outperformed the standard method when kinases and folate

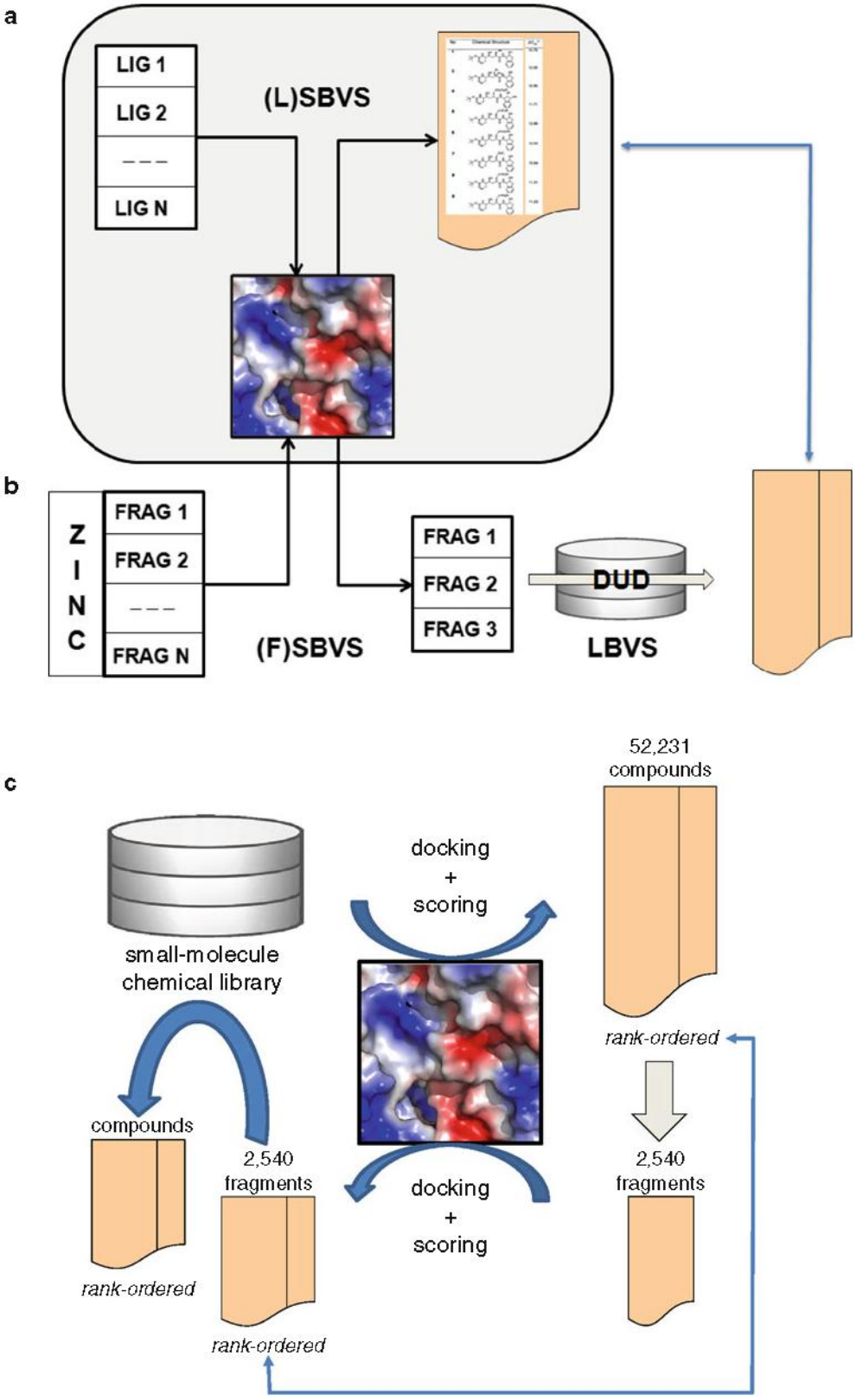


Fig. 1 Schematic representation of the classical ligand based (**a**) and fragment-based (**b**) SBVS procedures and the novel *contraction/selection/expansion* approach (**c**). The *blue double arrowheaded lines* stand for comparisons of the results between two given methods

enzymes were used as targets whereas true binders were recovered at similar rates in metalloenzymes and serine proteases. Strikingly, the worst performance for the new method was observed in the case of nuclear hormone receptors (NHR). When we analyzed the distribution of similarity indices between the ZINC fragments and the true

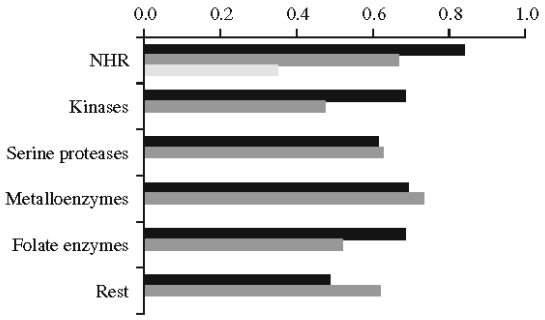


Fig. 2 AUCs obtained for the targets grouped into families. NHR: AR, ER_{ago}, ER_{antago}, GR, MR, PPARγ, PR, and RXRα; kinases: CDK2, EGFR, FGFR1, HSP90, P38 MAP, PDGFRB, SRC, TK, and VEGFR2; serine proteases: FXa, Thr, and trypsin; metalloenzymes: ACE, ADA, COMT, and PDE5; folate enzymes: DHFR and GART; and the rest: AChE, ALR2, AmpC, COX-1, COX-2, GPB, HIVPR, HIVRT, HMGR, INHA, NA, PARP, PNP, and SAHH (for the abbreviations see Table 1). *Black* and *grey bars* correspond to the fragment-based protocol presented here and the standard SBVS protocol with full ligands, respectively. An additional *bar* (*light gray*) was included in the case of NHR targets to highlight the performance of the new approach when the original collection of fragments was used

binders for these targets we realized that the latter were not properly represented using the original fragment collection, as illustrated in Fig. 3 for two prototypical targets of this class: MR, and RXRα.

Upon inclusion of appropriate representative ad hoc fragments (Fig. 4) we obtained AUC values of 0.82 and 0.96, respectively, for these two NHR that compared very favorably to the previous 0.13 and 0.24 that were achieved when these fragments were not included, and also to the 0.78 and 0.92 obtained with the standard protocol. In fact, the overall performance for the whole NHR family improved notably, as seen in Fig. 2 and also in Table 1, which displays the results upon incorporation of these new fragments. In view of this result, we tested the performance of fragments specifically derived from ligands bound to representative DUD targets (1/family) and found AUC values (0.78 for HSP90, 0.68 for ACE, 0.73 for AChE, and 0.74 for fXa) similar to those reported in Table 1.

These findings highlight the fact that as long as the fragment database is able to adequately represent the chemical diversity of the true binders, the present method can clearly outperform the standard classical SBVS procedure employing whole ligands. To better appreciate the structural similarities between true binders and decoys Fig. 5 shows the chemical structure of a query molecule (at the center), three true binders retrieved with high Tcs (at the bottom) and three decoys with low Tcs (at the top). The true binders's substructures that resemble the query fragment are highlighted in blue.

A fragment-based approach displaying similarities with our own has been published recently [25], although it appears to be more focused on the optimization of query molecules because the ligand database is decomposed into fragments that are evaluated for binding affinity using docking and scoring. Thereafter, those fragments exhibiting the lowest affinities are replaced by new ones and the affinity is re-calculated. The outcome is an optimized ligand made up of the best fragments. A more elaborate approach [26, 27] extends the query beyond the fragment itself by considering its microenvironment, which includes the relevant interacting protein residues. By compiling a diverse set of micro-environments (e.g. from the PDB) it is possible to optimize already known structures and/or suggest novel and improved compounds.

Comparative test of chemical diversity

Finally, to test the extent to which the chemical space represented by a set of docked compounds is covered by their corresponding fragments once they have been docked

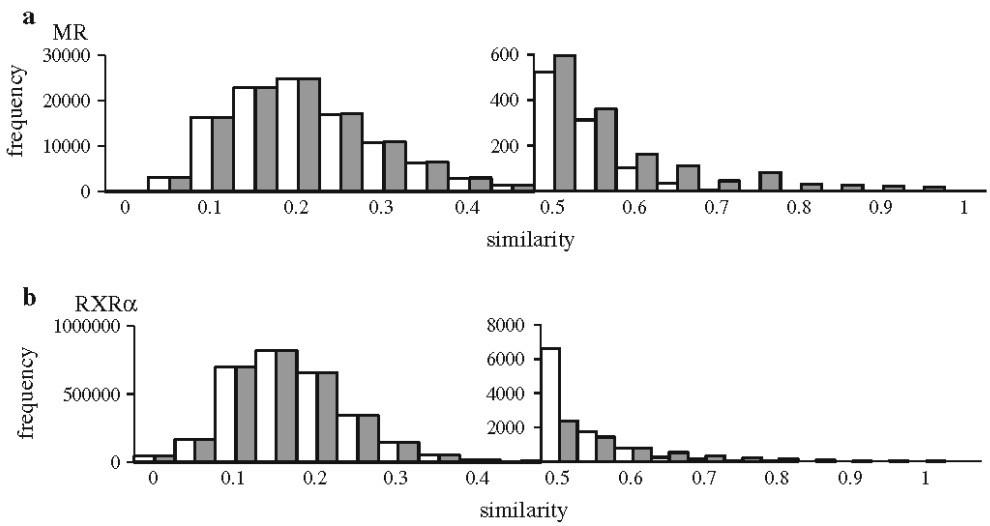


Fig. 3 *Histograms* showing the distribution of similarity indices between the ZINC fragments and the ligands (true binders + decoys) for two NHR targets (for the abbreviations see Table 1), MR (**a**) and RXRα (**b**), before (*white bars*) and after (*grey bars*) incorporation of the new fragments. Note that for Tc values (*x*-axes) higher than 0.5 the *y*-axes have been *scaled* up to *highlight* that the main differences occur in this region

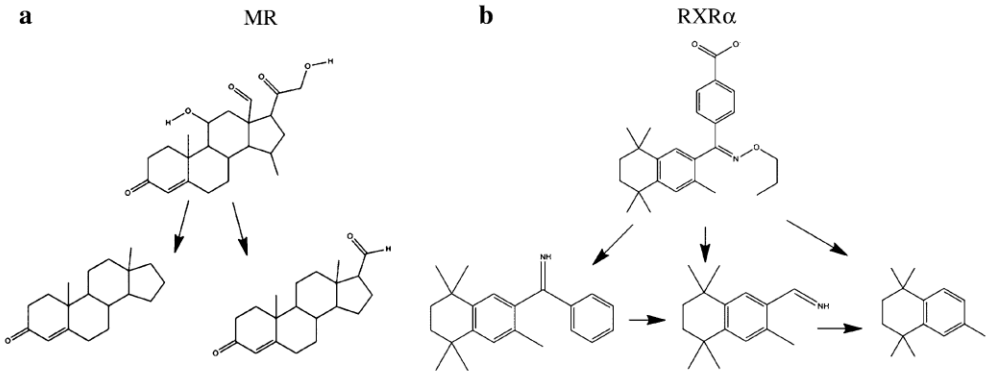


Fig. 4 MR (**a**) and RXRα (**b**) true binders and their appropriate fragment decomposition. These fragments were not generated in the default procedure that yielded the original collection of fragments

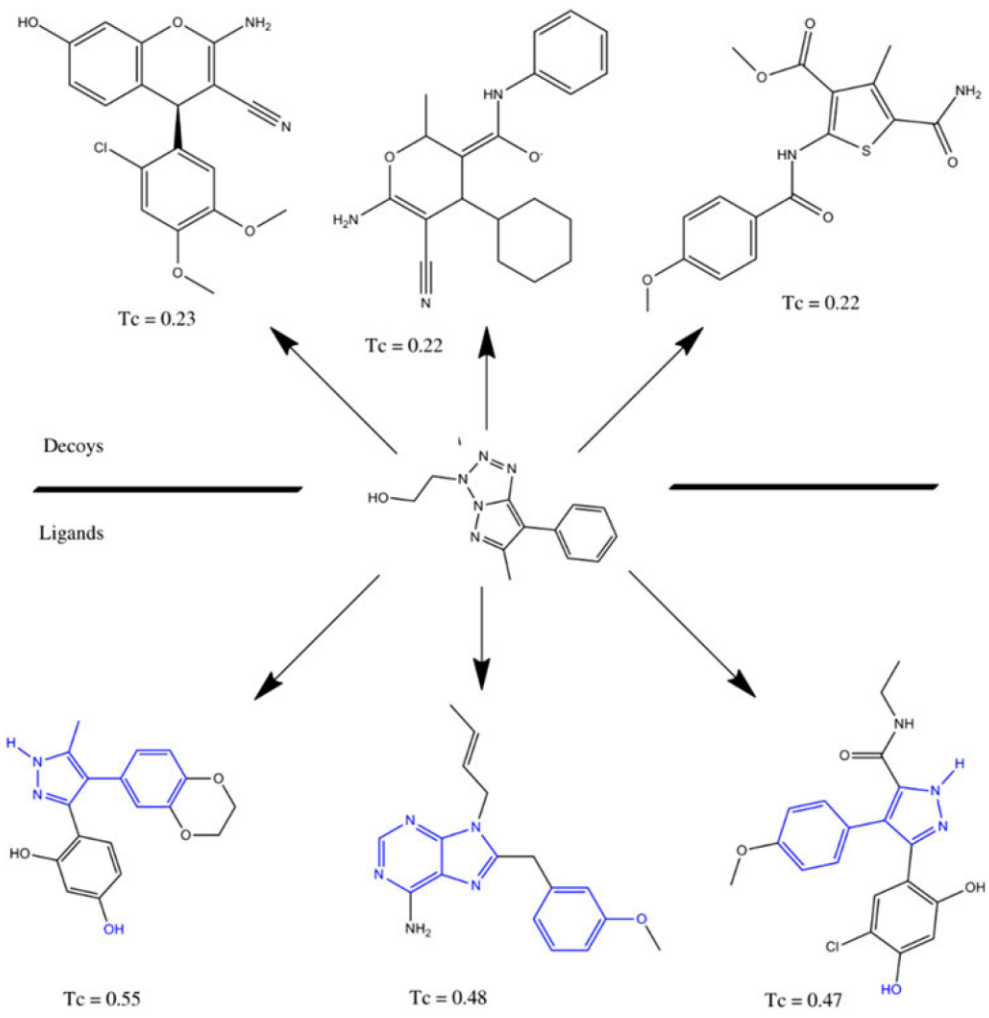


Fig. 5 Structural similarities between a query molecule (at the *center*), three true HSP90 binders (at the *bottom*), and three decoys (at the *top*). The true binders' substructures contained in the query are *highlighted* in *blue*

into the same target, we took 52,231 molecules from an in-house chemical library that had been subjected to SBVS and ranked by docking scores. The all-versus-all comparison between the 2,540 non-redundant fragments and the 52,231 parent compounds afforded a modest value (50%) for the overlap (*O*). This percentage reflects that considerable noise is introduced when the comparison is made without a previous filtering step, that is, the overlap one might expect when a brute force approach is employed. This value, however, is increased to 72% when the top 1% compounds from both rank-ordered lists are considered, which suggests that a greater success can be achieved when compound selection is made on the basis of a more robust criterion such as the complementarity between the fragments and the binding site, as represented here by the docking step. Furthermore, when only the 3 best-scoring fragments were selected for the comparison, an overlap of 82% was obtained. This indicates, again, that careful

fragment selection is likely to result in rather successful solutions that efficiently cover the chemical space explored by a traditional SBVS protocol. Therefore we believe that further improvements on this novel approach can be expected by refining the fragment generation procedure and by fine-tuning the similarity searches.

Benchmarking

Our in-house VS platform VSDMIP [28] was used for all the calculations. The input is a SMILES string representation for each ligand or fragment. These SMILES are inserted into the VSDMIP database by processing them as outlined before (conversion to 3D, charges and radii assignment, and conformational analysis). On average, using either a 32-bit 3.2 GHz PIV or a 32-bit 3.06 GHz Xeon processor, VSDMIP is able to insert into the database 2,600 fragments per CPU and day. This is roughly the same number of ligands that can be typically processed. Taking into account that fragments, by definition, are smaller in size than ligands, the equivalence in these insertion times reflects the structural complexity of the fragments. In fact, they are not fragments in the strictest sense as they have not been generated from splitting ligands into smaller pieces. Rather, they are small ligands that fulfill the properties arbitrarily defined for fragments. On the other hand, in terms of docking, and due to their reduced size compared to typical ligands, processing times are greatly shortened (248 ligands vs. 553 fragments docked per CPU and day, using the same CPUs as above). Finally, VSDMIP can perform 2.4×10^9 comparisons/day and CPU using 2D fingerprints. With all these numbers in hand, it can be stated that they are affordable on almost any small- to medium-size cluster ($\approx$25–100 processors) and would allow users to perform several complete protocols (as the one described here)/day. In this way, many tests can be easily conducted, which increases the likelihood of obtaining new promising hits.

Conclusions

We have presented a new way to combine SBVS and LBVS strategies that consists of three steps: (1) organize and reduce a ligand database into a non-redundant fragment library (*contraction* step), (2) dock all of these fragments into the binding pocket of the macromolecular target and select the three best-scoring ones (*selection* step, SBVS), and (3) perform a similarity analysis using the selected fragments to interrogate the database and select the most similar ligands (*expansion* step, LBVS). Compared to a typical SBVS campaign, computer running times are considerably reduced as a consequence of the fragments' smaller size relative to typical ligands and their non-redundancy. On the other hand, chemical space coverage is not critically compromised. This protocol allows the user to focus on specific regions within the chemical space present in the ligand database by enabling: (1) more efficient VS runs, (2) development of focused virtual libraries, and (3) scaffold hopping for chemical modification. Provided that the fragments have been previously obtained (e.g. using CDK as mentioned above), all the operations can be performed within the VSDMIP graphical environment, which was implemented as a PyMOL plugin [28]. Finally, it should be mentioned that the performance of the method is expected to be highly dependent on the ligand database that is employed to obtain the fragments: the greater the diversity, the better the coverage of chemical space.

Acknowledgments This work was supported by grants from Ministerio de Ciencia e Innovación (MICINN) BIO2008-04384 (to Antonio Morreale) and SAF2009-13914-C02-02 (to Federico Gago), and Comunidad Autónoma de Madrid (CAM) S-BIO-0214-2006 (BIPEDD) and S2010-BMD-2457 (BIPEDD-2). Antonio Morreale acknowledges CAM for financial support to the Fundación Severo Ochoa through the AMAROUTO program. Álvaro Cortés-Cabrera thanks Ministerio de Educación for the FPU Grant AP2009-0203. We are grateful to OpenEye Scientific Software, Inc. for providing us with an academic license for their software. The technical support and advice from the Bioinformatics team at CBMSO is gratefully acknowledged.

References

1. Shuker SB, Hajduk PJ, Meadows RP, Fesik SW (1996) Discovering high-affinity ligands for proteins: SAR by NMR. Science 274(5292):1531–1534
2. Hajduk PJ, Greer J (2007) A decade of fragment-based drug design: strategic advances and lessons learned. Nat Rev Drug Discov 6(3):211–219. doi:10.1038/nrd2220
3. Lipinski C, Hopkins A (2004) Navigating chemical space for biology and medicine. Nature 432(7019):855–861. doi:10.1038/nature03193
4. Hann MM, Leach AR, Harper G (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. J Chem Inf Comput Sci 41(3):856–864. doi:10.1021/ci000403i
5. Schuffenhauer A, Ruedisser S, Marzinzik AL, Jahnke W, Blommers M, Selzer P, Jacoby E (2005) Library design for fragment based screening. Curr Top Med Chem 5(8):751–762
6. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. J Chem Inf Model 45(1):177–182. doi:10.1021/ci049714+
7. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46(1–3):3–26
8. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49(23):6789–6801. doi:10.1021/jm0608356
9. Perez C, Ortiz AR (2001) Evaluation of docking functions for protein-ligand docking. J Med Chem 44(23):3768–3785

10. PDB IDs 1GPK, 1HWW, 1IA1, 1JD0, 1J3J, 1HNN, 1HQ2, 1IG3, 1K3V, 1LRH, 1N1M, 1N2V, 1OF1, 1OF6, 1OWE, 1P2Y, 1P62, 1Q1G, 1Q41, 1Q4G, 1R9O, 1SG0, 1SQW, 1TOW, 1TT1, 1TZ8, 1UIC, 1U4D, 1UOU, 1W1P, 1W2G, 1X8X, and 1XM6
11. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WT, Mortenson PN, Murray CW (2007) Diverse, high-quality test set for the validation of protein-ligand docking performance. J Med Chem 50(4):726–741. doi:10.1021/jm061277y
12. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28(1):31–36. doi:10.1021/ci00057a005
13. Open Babel (2011) The open source chemistry toolbox. http://openbabel.org/wiki/Main_Page. Accessed 01 March 2012
14. Corina (2000) Computerchemie Langemarckplatz 1 E, Germany, MNG
15. Stewart JJP (1990) MOPAC: a semiempirical molecular orbital program. J Comput Aided Mol Des 4(1):1–103. doi:10.1007/bf00128336
16. Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ (2005) The Amber biomolecular simulation programs. J Comput Chem 26(16):1668–1688. doi:10.1002/jcc.20290
17. Gil-Redondo R (2006) Master thesis UNED, Madrid
18. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA (2004) PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. Nucleic Acids Res 32(Web Server issue):W665–W667. doi:10.1093/nar/gkh381
19. Murray CW, Baxter CA, Frenkel AD (1999) The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. J Comput Aided Mol Des 13(6):547–562
20. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) The chemistry development kit (CDK): an open-source Java library for chemo- and bio-informatics. J Chem Inf Comput Sci 43(2):493–500. doi:10.1021/ci025584y
21. Voigt JH, Bienfait B, Wang S, Nicklaus MC (2001) Comparison of the NCI Open Database with Seven Large Chemical Structural Databases. J Chem Inf Comput Sci 41(3):702–712. doi:10.1021/ci000150t
22. Schneider G, Neidhart W, Giller T, Schmid G (1999) "Scaffold–Hopping" by topological pharmacophore search: a contribution to virtual screening. Angew Chem Int Ed Engl 38(19):2894–2896
23. Sandor M, Kiss R, Keseru GM (2010) Virtual fragment docking by Glide: a validation study on 190 protein-fragment complexes. J Chem Inf Model 50(6):1165–1172. doi:10.1021/ci1000407
24. Verdonk ML, Giangreco I, Hall RJ, Korb O, Mortenson PN, Murray CW (2011) Docking performance of fragments and druglike compounds. J Med Chem 54(15):5422–5431. doi:10.1021/jm200558u
25. Lin FY, Tseng YJ (2011) Structure-based fragment hopping for lead optimization using predocked fragment database. J Chem Inf Model 51(7):1703–1715. doi:10.1021/ci200136j
26. Moriaud F, Doppelt-Azeroual O, Martin L, Oguievetskaia K, Koch K, Vorotyntsev A, Adcock SA, Delfaud F (2009) Computational fragment-based approach at PDB scale by protein local similarity. J Chem Inf Model 49(2):280–294. doi:10.1021/ci8003094
27. Durrant JD, Friedman AJ, McCammon JA (2011) CrystalDock: a novel approach to fragment-based drug design. J Chem Inf Model 51(10):2573–2580. doi:10.1021/ci200357y
28. Cabrera AC, Gil-Redondo R, Perona A, Gago F, Morreale A (2011) VSDMIP 1.5: an automated structure- and ligand-based virtual screening platform with a PyMOL graphical user interface. J Comput Aided Mol Des 25(9):813–824. doi:10.1007/s10822-011-9465-6

Supplementary Information

A reverse combination of structure-based and ligand-based strategies for virtual screening

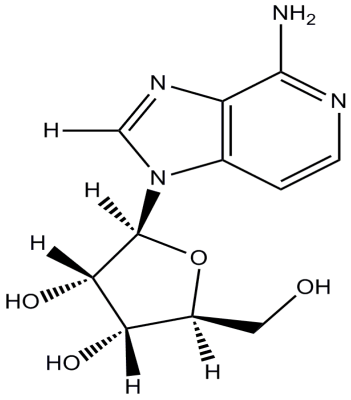
Álvaro Cortés-Cabrera[1,2], Federico Gago[1], and Antonio Morreale[2*]

[1]Departamento de Farmacología, Universidad de Alcalá, E-28871 Alcalá de Henares, Madrid, Spain.

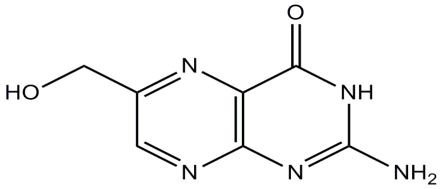
[2]Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC/UAM), Campus UAM, c/ Nicolás Cabrera 1, E-28049 Madrid, Spain.

*__Corresponding author:__ Antonio Morreale. Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC/UAM), Campus de Cantoblanco, c/ Nicolás Cabrera 1, E-28049 Madrid, Spain. E-mail: amorreale@cbm.uam.es. Telephone number: + 34 911 964 633. Fax number: + 34 911 964 422.

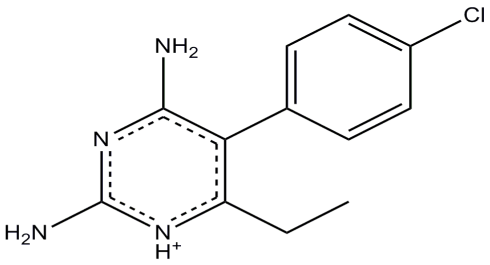
Fig. S1. Chemical structures and PDBIDs of the ASTEX ligands for which tautomers and protonation state were assigned manually.



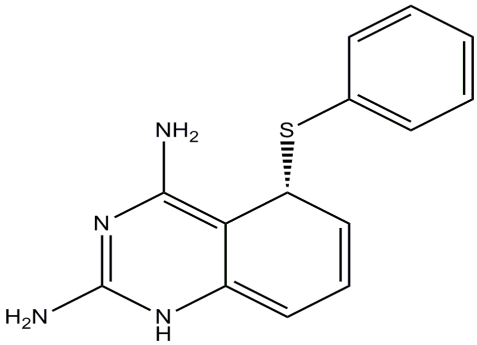
1hp0



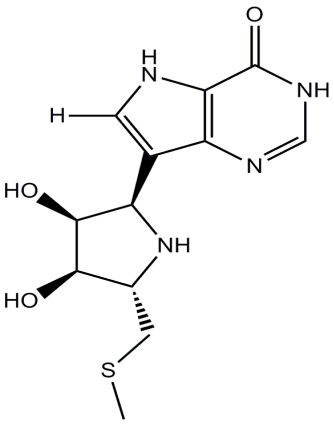
1hq2



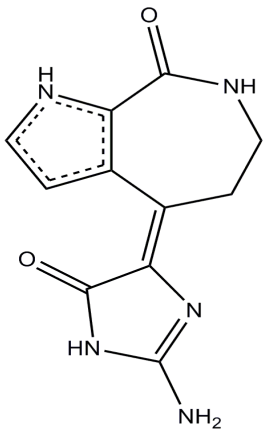
1j3j



1ia1



1q1g



1u4d

Table S1. Additional molecules (water, ions, and groups) added to the ASTEX targets.

Target	Additional molecules
1gpk	Water 2529
1hp0	Calcium ion 328
1ig3	Sulfate group 609 and water 631
1jd0	Zinc ion 901
1owe	Sulfate group 302
1p2y	Heme group 430
1q41	Water 630
1tow	Water 571
1w1p	Water 2102
1w2g	Water 2073
1x8x	Waters 648 and 869

Article IV

Comparison of Ultra-fast 2D and 3D Ligand and Target Descriptors for Side Effect Prediction and Network Analysis in Polypharmacology

Background and author's contribution

The advent of structure-based drug discovery provided the scientific community with a very reliable and rational model under the paradigm of one-illness one-target one-compound. On the contrary, previous discovery efforts were based on whole tissues or animals for testing the compounds. The main advantage of the new paradigm is the simplification of the tests and the confidence that the pharmacological effect is a direct consequence of the interaction of the drug with the proposed target. However, the downside includes three major caveats arising from the oversimplification of the model: i) toxicity, ii) metabolism and iii) lack of pharmacological effect.

The toxicity problem could be related to metabolism but, in most cases, it can be traced down to hitting secondary targets to which the compound was not intended to bind. The reduction of the model to just one target is more prone to this kind of problems since no other targets are taken into account in the early stages of development.

In silico metabolism prediction is an active field of research where most of the efforts are centered on creating models that rely on large databases of compounds for which several physicochemical properties have been measured. Nevertheless, other models based on physical principles or chemical reactivity do exist. These models have had moderate success in the prediction of which compounds are more likely to be substrates for hepatic enzymes or to be absorbed erratically.

Finally, sometimes compounds that do bind to a certain target, which is related to a pathological process, may not show any therapeutic effect. In addition, it is known that several drugs on the market (mostly for psychiatric disorders) have a great impact on the illness due to the modulation of many targets simultaneously. These effects are hardly considered under the one-illness one-target paradigm and are also difficult to address from the computational standpoint where problems exist even when only the interaction of one drug with one target is modelled.

Lately a comeback of the so-called polypharmacology has surged under the new paradigm of *systems pharmacology*. In this regard, this *new* field tries to address the caveats of the current model borrowing concepts and ideas from systems biology with an emphasis on network-like schemes where relations amongst targets and compounds are taken into account during the drug discovery process.

In the following work, we analyzed the relationships and information present in networks built from comparing targets and ligands, and we tried to provide a simple tool to investigate these relationships and to predict possible targets and side effects for a given compound.

The author of the manuscript designed and developed the tools and methodologies described in the following document, performed the analyses and wrote the initial version of the paper.

BJP British Journal of Pharmacology

DOI:10.1111/bph.12294
www.brjpharmacol.org

RESEARCH PAPER

Comparison of ultra-fast 2D and 3D ligand and target descriptors for side effect prediction and network analysis in polypharmacology

Alvaro Cortés-Cabrera[1,2], Garrett M Morris[3,4], Paul W Finn[3], Antonio Morreale[1,*] and Federico Gago[2]

[1]*Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC/UAM), Madrid, Spain,* [2]*Departamento de Ciencias Biomédicas, Universidad de Alcalá, Madrid, Spain,* [3]*InhibOx Ltd, Oxford Centre for Innovation, Oxford, UK, and* [4]*Crysalin Ltd, Cherwell Innovation Center, Oxfordshire, UK*

Correspondence
Professor Federico Gago, Department of Biomedical Sciences, Universidad de Alcala, Alcala de Henares, E-28871 Madrid, Spain. E-mail: federico.gago@uah.es

*Present address: Repsol Technology Center, E-28923 Móstoles, Madrid, Spain.

Keywords
adverse drug reactions; chemical fingerprints; drug targets; polypharmacology; side effects

Received
10 April 2013
Revised
24 June 2013
Accepted
2 July 2013

BACKGROUND AND PURPOSE

Some existing computational methods are used to infer protein targets of small molecules and can therefore be used to find new targets for existing drugs, with the goals of re-directing the molecule towards a different therapeutic purpose or explaining off-target effects due to multiple targeting. Inherent limitations, however, arise from the fact that chemical analogy is calculated on the basis of common frameworks or scaffolds and also because target information is neglected. The method we present addresses these issues by taking into account 3D information from both the ligand and the target.

EXPERIMENTAL APPROACH

ElectroShape is an established method for ultra-fast comparison of the shapes and charge distributions of ligands that is validated here for prediction of on-target activities, off-target profiles and adverse effects of drugs and drug-like molecules taken from the DrugBank database.

KEY RESULTS

The method is shown to predict polypharmacology profiles and relate targets from two complementary viewpoints (ligand- and target-based networks).

CONCLUSIONS AND IMPLICATIONS

The open-access web tool presented here (http://ub.cbm.uam.es/chemogenomics/) allows interactive navigation in a unified 'pharmacological space' from the viewpoints of both ligands and targets. It also enables prediction of pharmacological profiles, including likely side effects, for new compounds. We hope this web interface will help many pharmacologists to become aware of this new paradigm (up to now mostly used in the realm of the so-called 'chemical biology') and encourage its use with a view to revealing 'hidden' relationships between new and existing compounds and pharmacologically relevant targets.

Abbreviations

ADR, adverse drug reactions; DUD, Directory of Useful Decoys; EVD, extreme value distribution; LBVS, ligand-based virtual screening; MDDR, MDL Drug Data Report; MMFF94, Merck Molecular Force Field; NHR, nuclear hormone receptors; PDB, Protein Data Bank; ROCS, Rapid Overlay of Chemical Structures; SEA, Similarity Ensemble Approach; SMILES, simplified molecular-input line-entry specification; WOMBAT, WOrld of Molecular BioAcTivity

Introduction

The famous 'magic bullet' term coined by Paul Ehrlich more than one hundred years ago in the field of chemotherapy (Ehrlich, 1907; Witkop, 1999) paved the way to the classical one-compound–one-target paradigm that has largely dominated drug discovery for the last 25 years or so. This reductionist concept has led to a limited appraisal of the causes underlying the side effects of commercial drugs, derived from modulation of secondary targets that can nevertheless play a fundamental role in explaining pharmacological profiles. This is particularly true for drugs acting on the CNS, which can bind to many different receptors ('magic shotgun'; Roth *et al.*, 2004), and for some multi-kinase inhibitors in oncology (Knight *et al.*, 2010). This realization suggests that a more direct approach to polypharmacology should be taken in modern drug discovery from the very early stages of screening and lead identification. A multi-target–multi-compound approach would provide a much more accurate description of the underlying pharmacology but, given the large size of both chemical and biological spaces, it is also harder to understand, hence the need for tailor-made computer programs that can handle and relate the enormous, and still increasing, amounts of bioactivity data available for both compounds and targets.

Network analysis (Hopkins, 2008; Berger and Iyengar, 2009) in systems pharmacology (Van Der Greef and McBurney, 2005) has recently emerged and promises to revolutionize the field of drug discovery. Polypharmacology, drug repurposing, target fishing and adverse effect prediction are some of the major applications made possible by this paradigm shift, which contrasts with the traditional one-ligand–one-target approach that is still in use in most high-throughput experimental and virtual screening campaigns nowadays (Ripphausen *et al.*, 2010). Traditionally, *in silico* target-fishing methods have been related to reverse docking (Cai *et al.*, 2006) using one or several compounds against multiple putative targets (Simon *et al.*, 2012). In recent years, however, ligand-based methods exploiting either fingerprints containing two- (2D) and three-dimensional (3D) chemical information or 3D shape descriptors (Bender *et al.*, 2007; Keiser *et al.*, 2007; Vidal and Mestres, 2010; Besnard *et al.*, 2012) have been employed to predict activity profiles and target–target relationships. Superpositional methods that either compare the shapes of two molecules by analytically optimizing their volume intersection (Grant *et al.*, 1996), as implemented in the program ROCS (Rapid Overlay of Chemical Structures, OpenEye Scientific Software, 2011; Rush *et al.*, 2005) or use a surface-based morphological similarity function while minimizing the overall molecular volume of the aligned structures, such as Surflex-Sim (Jain, 2000), have been shown to perform well in the task of predicting off-target activities of ligands (Ballester *et al.*, 2009; Yera *et al.*, 2011). These approaches, although successful, rely on direct pairwise comparisons and this can reduce global performance when database size grows above tens of millions of molecules (Wang *et al.*, 2009).

ElectroShape is a non-superpositional method for ultrafast comparison of ligands that expands the capabilities and improves the performance of the Ultra Shape Recognition methodology (Ballester and Richards, 2007) by incorporating the molecular charge distribution (Armstrong *et al.*, 2010). In brief, ElectroShape uses three spatial dimensions and adds partial charge as a fourth dimension to capture electrostatic information in the form of 15 descriptors that account for the first, second and third moments of the distributions of distances from five distinct points of the molecule (centroids) in a four-dimensional (4D) space. Molecular similarity is then calculated as the Manhattan distance between ElectroShape descriptors belonging to two different molecules. The following facts are key to increasing the speed of calculation and improving performance: (i) the descriptors are very small and can be pre-calculated and stored for each compound from ensembles of low-energy conformers, and (ii) the similarity calculation is non-superpositional and requires only a few mathematical operations. The value of Ultra Shape Recognition in ligand-based virtual screening (LBVS) has already been recognized (Ballester *et al.*, 2010).

In the following, we show how the ElectroShape method has been validated for off-target prediction ('target fishing'), LBVS and chemogenomic network analysis (Figure 1). The results obtained have been compared, firstly, to those produced by other well-known 2D techniques that make use of ligand and target annotations from bioactivity databases, and then to structural information from a database of ligand-binding sites in proteins. Implementation on an unsophisticated web server is also presented that can enable pharmacologists and other interested researchers to predict possible adverse effects and secondary targets for a given drug and to explore pharmacological space from the viewpoints of both ligand and target simultaneously.

Methods

Chemogenomic datasets

To explore the potential use of the ElectroShape approach (Armstrong *et al.*, 2010) for target fishing and prediction of adverse effects, and to compare it with existing 2D methods (Hert *et al.*, 2008), we first built a chemogenomic database using information available from DrugBank (Knox *et al.*, 2011). To this end, target and drug sets were downloaded from http://www.drugbank.ca/ and imported locally into a MySQL relational database. Molecules were then parsed using RDKit (Landrum, 2011) to generate canonical simplified molecular-input line-entry specification (SMILES) strings (Weininger, 1988). Drug–target associations were also imported from DrugBank using Python scripts (http://www.python.org/).

For the 2D analysis, Morgan fingerprints, which are roughly equivalent to the Extended-Connectivity Fingerprints (ECFP4) commonly used in other target-fishing applications (Hert *et al.*, 2008; Rognan and Meslamani, 2011), were generated using RDKit with a radius of two bonds and 2048 bits and inserted in the database. A simple in-house chemistry cartridge was used to allow searching within the database (Cabrera *et al.*, 2011).

To perform the 4D ElectroShape analysis (Cartesian coordinates and charges), the SMILES string representations of the drugs were converted into 3D structures using CORINA (Sadowski *et al.*, 1994; 2003) and point charges were assigned

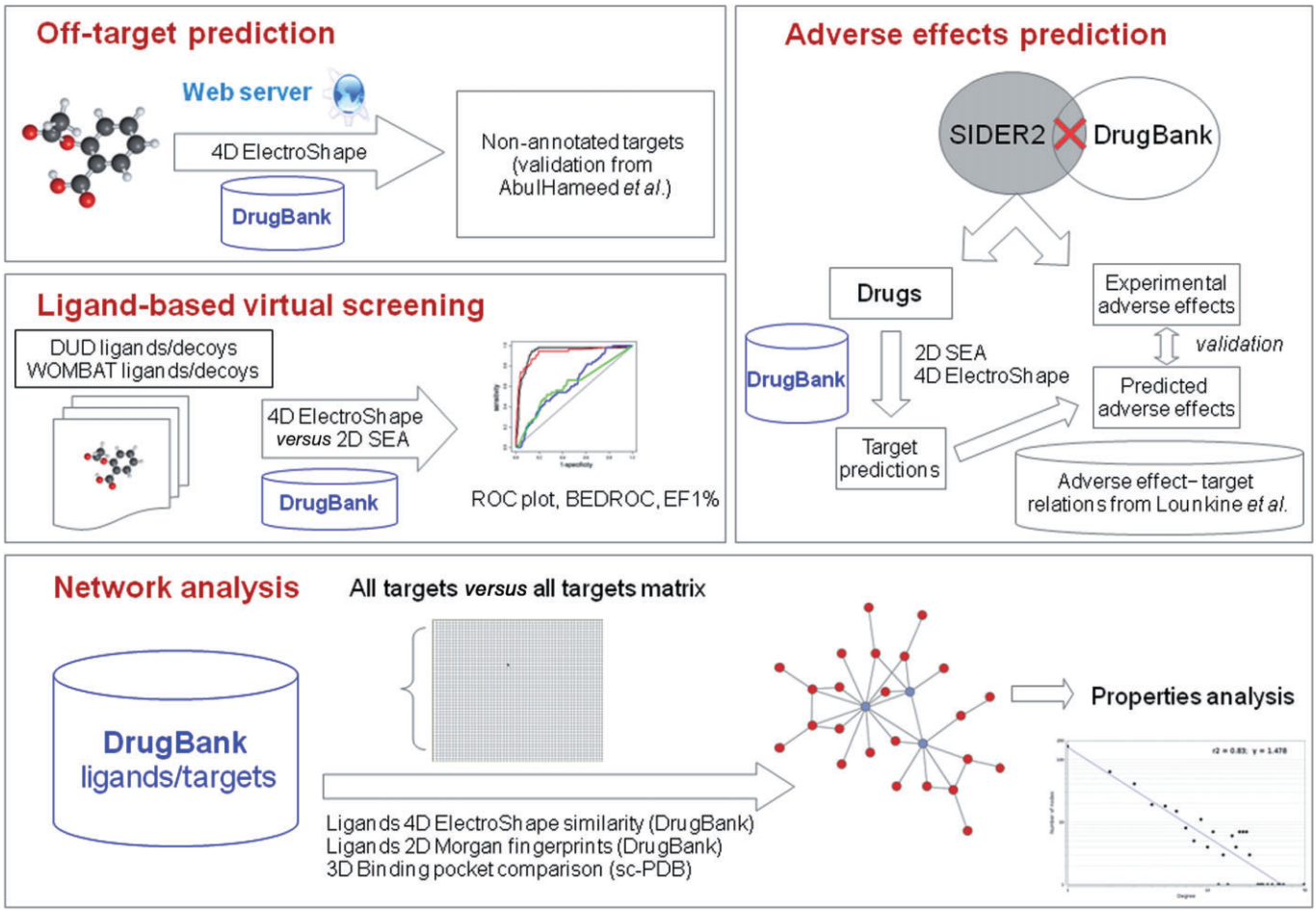


Figure 1

Scheme depicting the workflow presented in this article (see main text for details).

by OpenBabel making use of the Merck Molecular Force Field (MMFF94; Halgren, 1996). Then, a conformational analysis using ALFA (Cabrera *et al.*, 2011), a rule-based conformer generator similar to OMEGA (Hawkins *et al.*, 2010), was undertaken setting the maximum number of conformers to 200 and activating the maximum dissimilarity option to generate the most diverse ensemble possible. By selecting this number we ensure that most of the conformational space is covered for each ligand and avoid the risk of losing alternative similarity matches, a possibility that is almost a certainty when just a single low-energy conformation is taken as a representative 3D structure of ligands with a high number of rotatable bonds. Finally, the molecules were converted to Protein Data Bank (PDB) format (Berman *et al.*, 2007) using OpenBabel (O'Boyle *et al.*, 2011) and 4D descriptors were generated by ElectroShape (Armstrong *et al.*, 2010) for each molecular conformer and inserted into the database to allow simplified queries within chemical space using the cartridge.

To explore the similarity between targets in biological space, we used sc-PDB (Meslamani *et al.*, 2011), an annotated database that contains 3D coordinates for bound ligands and ligand-binding sites, as extracted from the PDB. The sc-PDB complexes were filtered so that only those with a direct match in the DrugBank set were kept.

Similarity calculation and drug–drug and target–target matrices

To calculate target similarity from 2D ligand descriptors, we used the Similarity Ensemble Approach (SEA; Keiser *et al.*, 2007) and Morgan fingerprints (Landrum, 2011) of all molecules, which are annotated to have an effect on every DrugBank target. To this end, an all-drug versus all-drug matrix was built by comparing the fingerprints using the Tanimoto coefficient (Rogers and Tanimoto, 1960). Then, the scores of all the drugs associated with a given target were summed and a final *Z*-score for each target pair was computed. To obtain the *Z*-score for the similarity between target A and target B, it is necessary to calculate the expected random average (μ) and standard deviation (σ) of the total score of any two ensembles of $n \times m$ drugs (n being the number of drugs that bind to target A and m the number of drugs that bind to target B). Drugs present in DrugBank were randomly grouped in several ensembles ranging in size from 1 to 50 molecules (which is the common range of drugs per target found in DrugBank) and the total score was computed. Then, these scores were fitted to a power law equation and used to calculate the expected values for random targets and the corresponding *Z*-scores:

$$Z\text{-score} = (\Sigma \text{ scores} - \mu)/\sigma$$

In the case of ElectroShape's 4D ligand similarity, we followed an analogous method. Similarities between molecules were computed by calculating the Manhattan distance between each pair of conformer's descriptors for both molecules and the maximum value was kept as the final score between the two ligands. Individual scores between each pair of molecules were summed and the *Z*-score was calculated using the same procedure as for 2D comparisons: generation of random sets of different sizes followed by calculation of the best power law fit for both average and standard deviation of the total sum score.

Finally, for ligand-binding site similarity, we first used the protein structure alignment algorithm TM-align (Zhang and Skolnick, 2005) to align the binding sites of the targets present in both DrugBank and sc-PDB and then a template modelling score (TM-score) was extracted as a normalized measure of the topological similarity between any two target binding sites.

Web server implementation

The ElectroShape Polypharmacology web server was implemented in Python using the Django Web framework (https://www.djangoproject.com/). It is based on a MySQL database extended with a cartridge previously developed by the authors (Cabrera *et al.*, 2011). The user only has to paste a SMILES string into the search box and then select a method for target fishing. If the traditional 2D SEA method is chosen, the SMILES string is used to generate Morgan fingerprints and perform the query to the database. If an ElectroShape-based method is requested, the server (i) converts this string into a 3D structure, (ii) explores the conformational space using the rule-based program ALFA (Cabrera *et al.*, 2011), (iii) assigns MMFF94 atom point charges using OpenBabel and (iv) calculates the ElectroShape 4D descriptors.

The network representation is based on the Javascript InfoVis toolkit and the information is extracted directly from the database through Django. Finally, 2D images for the small molecules are generated on-the-fly using the RDKit module from Django.

Prediction of side effects and off targets

Reported associations extracted from the literature (Lounkine *et al.*, 2012) between pharmacological effects (including those categorized as 'adverse effects') and certain targets were imported into a MySQL relational database. On the other hand, adverse drug reactions (ADR) and corresponding drugs' SMILES strings were downloaded from the side effect resource SIDER 2 (Kuhn *et al.*, 2010) and a subset comprising 151 drugs not present in DrugBank was created and used for validating our results. For each of these drugs, we calculated the putative binding profile and queried the database for the adverse effects associated to the predicted targets. The goodness of our predictions was assessed by evaluating how many ADR are correctly identified and/or missed out of the total number of 270 ADR contained in the MySQL database. This problem was handled as a group of 270 binary decisions for each drug: 1 = true association; 0 = no association. The true positive and negative rates, as well as the false positive and negative rates, were calculated for each drug and averaged.

For prediction of non-annotated off-targets, the dataset proposed by (AbdulHameed *et al.*, 2012) was used. SMILES strings were extracted from PubChem when possible, or generated manually.

LBVS

To perform an on-target validation, we used the Directory of Useful Decoys (DUD) (Huang *et al.*, 2006) and the 'WOrld of Molecular BioAcTivity' (WOMBAT) (Good and Oprea, 2008) subset to account for the analogue bias in the set of active compunds. Ligands and decoys were extracted from the original distributions and processed in an identical way to that followed with the DrugBank molecules in order to allow direct search and comparison because ElectroShape results are dependent on the partial charge method used. Briefly, the original files in multi-molecule structure data format were used to generate directly Morgan fingerprints and SMILES strings using RDKit. The latter were employed to generate initial 3D structures using CORINA (Sadowski *et al.*, 2003) and different conformations for each of them were obtained by using ALFA (Cabrera *et al.*, 2011).

We were able to match every target in the DUD to a corresponding DrugBank target with at least one representative ligand except for β-lactamase AmpC whose only protein-bound ligand was not transformed properly, hence results are presented for only 39 targets. As a measure of the global similarity per ligand, we used the SEA *Z*-score for 2D and 4D methods, and in the case of ElectroShape 4D, we also used the maximum similarity values among the compounds per target so that comparison with other data fusion techniques could be performed.

Network analysis

The target–target matrix was analysed and the *Z*-score distribution was fitted to an extreme value distribution (EVD) in order to calculate the expectation values (*E*-values, e.g. a probability of observing a given *Z*-score using random data) for the similarity between targets (Hert *et al.*, 2008). Using different cut-off values, we transformed those matrices into threshold networks, adding an edge between two targets or nodes if the *E*-value or a similar score was above a certain value that depends on the kind of comparison being made (2D ligand chemistry, 3D ligand-binding site or 4D ligand shape and charge distribution). For the ligand-binding sites, we used the TM-score values to build the corresponding threshold network.

The Cytoscape software and network analysis plug-in (Smoot *et al.*, 2011) was employed to study several statistical properties and to compute the union, difference and intersection of 2D ligand, 4D ligand and 3D receptor networks. Finally, to explore the usefulness of the information in the networks, a 'triviality' test was performed for each kind of network. In this test, we counted the number of edges due to a name similarity (string matching over 0.6) and percentage of common compounds between targets of more than 60%. These indicators try to measure the quantity of information that could be extracted trivially from the names of the targets or, at plain sight, from the profiles of the ligands.

Results

Chemogenomic datasets

After applying to DrugBank the protocol described above, the database contained a total of 5685 molecules and 3779 targets related by 11 559 annotated activities. In the case of the 3D database for ElectroShape, we ended up with 565 505 different conformers. This number roughly corresponds to an average of 100 conformers per molecule. Only sc-PDB ligand-binding sites belonging to targets present in DrugBank were selected and identified by their UniProt (Wu *et al.*, 2006) identification code. This procedure yielded a total of 1056 targets.

Similarity calculation and drug–drug and target–target matrices

Inspired by the work from Shoichet's lab (Hert *et al.*, 2008) that related biological targets by employing the 2D chemical descriptors of their ligands, we analogously developed a parameterization for the SEA (Keiser *et al.*, 2007) method using random groups of molecules from DrugBank. As can be seen in Table 1, after the fitting procedure, power law values are comparable to the original values obtained by Shoichet *et al.* (Hert *et al.*, 2008) for the MDL Drug Data Report (MDDR) database (MDDR, 2006). The SEA method was also re-parameterized using ElectroShape descriptors and an equivalent random set of molecules.

Finally, for the 3D receptor set arising from the intersection of DrugBank and sc-PDB databases, we obtained an all-against-all matrix for the 1056 targets using the TM-score normalized by the average number of residues of the two ligand-binding sites being compared.

Off-target validation results

Using a limited test set from DrugBank (see under Methods) that consisted of recently discovered but not yet annotated secondary targets for a certain number of compounds (Ballester *et al.*, 2009), we tested the ability of the ElectroShape chemogenomic approach to predict these annotations correctly. For three of the four compounds (Table 2), the primary activity was found to be ranked first in the target profile (except for RO-25–6981, in which case it was found in the third place), while all secondary targets were identified within the first seven predicted targets and with very close similarity values to the first scoring target.

Prediction of adverse effects

Following on recently published work (Lounkine *et al.*, 2012), we related certain targets belonging to the DrugBank set to 290 probable adverse effects. We then predicted the 2D SEA polypharmacology profiles of the drugs included in the SIDER 2 database but not in DrugBank, and we extracted the adverse effects related to those targets. For every molecule, the specificity and selectivity of the method were calculated

Table 1

Fitting comparison for DrugBank and Shoichet *et al.* sets in SEA

	Mean exponent	Mean coefficient	Pearson r^2	Standard deviation exponent	Standard deviation coefficient	Pearson r^2
DrugBank 2D Morgan fingerprints	1.01	5.81×10^{-3}	0.9991	0.534	6.03×10^{-2}	0.9977
DrugBank ElectroShape 4D	0.99	4.76×10^{-2}	0.9998	0.635	1.56×10^{-1}	0.9951
Shoichet *et al.* (Hert *et al.*, 2008)	1	4.24×10^{-4}	0.9998	0.665	4.49×10^{-3}	0.9882

Table 2

Off-target test set from (AbdulHameed *et al.*, 2012)

Compound	Primary target rank	Secondary target rank
Dimetholizine	1 (histamine H_1 receptor)	6 (α_1A adrenoceptor)
		7 (α_1D adrenoceptor)
		7 (α_1B adrenoceptor)
		7 (5-HT_1A receptor)
		1 (D_2 dopamine receptor)
Denopamine	1 (β_1 adrenoceptor)	3 (β_3 adrenoceptor)
Ifenprodil	1 (NMDA receptor)	3 (μ opioid receptor)
RO-25–6981 (NMDA receptor ligand)	3 (NMDA receptor)	4 (D_4 dopamine receptor)
		1 (noradrenaline transporter)
		1 (5-HT transporter)

and then the values were averaged, yielding 92% for specificity and 16% for selectivity. We also obtained a minimum of 20% of adverse effects predicted for 43% of the compounds. These values appear to indicate that most of the adverse effects that could be predicted are missing but the false positive rate is within a reasonable range. With ElectroShape, we obtained better values, namely, a 95% of specificity and 22% of selectivity, which highlights the improved ability of this method to discover additional targets over the general chemical scaffold comparison merely using 2D fingerprints.

LBVS

Regarding the on-target validation of the dataset (cf. Figure 1), we selected the DUD and a subset of WOMBAT to avoid, to some extent, the effect of analogues in the true binders sets. In the case of the 4D method, we also tested the SEA (4D SEA) and the maximum similarity method (4D MAX).

By-target and average results for AUC of receiver operating characteristic plots are given in Table 3. A perfectly accurate method would have an AUC of 1, while a random method would have an AUC of 0.5. It can be seen that the 2D and 4D MAX methods perform equally and that both outperform the 4D SEA. A closer look at the WOMBAT results (Table 4) reveals that a strong analogue bias is present for all methods but especially for 2D SEA, the performance of which drops more than those of the two 4D methods.

It is worth noting that the average performance of our ElectroShape method compares favourably with those obtained through other means. In particular, the 4D MAX approach compared favourably with ROCS max ComboScore (Ballester *et al.*, 2009; reported average of 0.77). Besides, the non-superpositional ElectroShape 4D method is several orders of magnitude faster than ROCS (Grant *et al.*, 1996) because (i) the molecular descriptors are pre-calculated, (ii) a direct comparison requires just a Manhattan distance calculation and (iii) ElectroShape, unlike ROCS, does not require a computationally expensive superposition of one ligand onto another to compute the optimal ligand similarity.

Network analysis

Based on the target–target similarity matrices, we built several threshold networks with different cut-off values and compared several properties. For the case of ligand-based networks, comparison between ligands yielded *Z*-scores, which are directly transformable into *E*-values through an EVD fitting procedure. In the case of structure-based networks, the direct comparison between ligand-binding sites yielded TM-score values, which were used as the threshold criteria to trace an edge between two nodes.

To examine other properties of the network, we computed the union, intersection and differences between the three kinds of networks (ligand-based 2D and 4D, and receptor-based 3D) and calculated the triviality score as the percentage of the network edges built from plain-sight knowledge deduced from ligand structures and target names.

According to the network statistical parameters (cluster coefficient, characteristic path length and node degree distribution, which accounts for the distribution of the number of neighbours per node, Figure 2), the ligand-based network could be classified as a broad-scale and small-world network. This is so because it presents a high cluster coefficient, a short characteristic path (e.g. nodes are reachable from others within a few leaps, resulting in the small-world property, Figure 2) and a power law fitting of the node degree distribution (broad scale). Remarkably, the same classification is also present in the 3D target network, whereas the properties of the PSI-BLAST-based (Altschul *et al.*, 1997) network presented by Shoichet *et al.* (Hert *et al.*, 2008) were not compatible with the categorization of broad-scale and small-world networks.

Ligand-based networks appear to be populated by very close clusters, which are themselves mostly unconnected and represent pharmacological families such as GPCR, kinases, metabolic enzymes, proteases and nuclear hormone receptors (NHR). This trend can also be observed from the 3D target network where some clusters become even clearer, that is, the NHR family because their members share very similar ligand-binding sites (Figure 3).

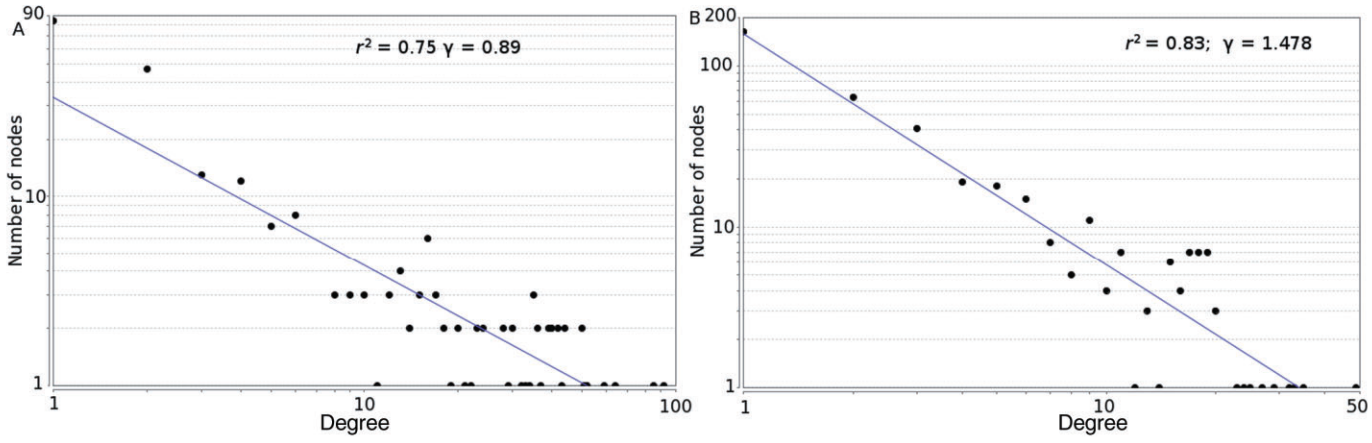


Figure 2

Node degree distribution for ElectroShape threshold network (A) and 3D binding site network (B).

Table 3

AUC results for the DUD using the three approaches tested

	4D SEA	4D MAX	2D SEA
ACE	0.59	0.71	0.81
AChE	0.57	0.61	0.64
ADA	0.59	0.64	0.91
ALR2	0.64	0.82	0.58
AmpC	–	–	0.55
AR	0.56	0.84	0.64
CDK2	0.55	0.78	0.72
COMT	0.54	0.63	0.80
COX1	0.55	0.74	0.69
COX2	0.66	0.91	0.46
DHFR	0.77	0.78	0.99
EGFR	0.77	0.73	0.95
ER_agonist	0.84	0.86	0.71
ER_antagonist	0.77	0.87	0.94
FGFR1	0.36	0.52	0.40
FXa	0.56	0.45	0.71
GART	0.93	0.92	0.77
GPb	0.70	0.81	0.86
GR	0.44	0.60	0.75
HIVPr	0.41	0.42	0.77
HIVRT	0.48	0.48	0.41
HMGA	0.86	0.96	0.97
HSP90	0.48	0.90	0.91
InhA	0.38	0.64	0.62
MR	0.69	0.84	0.88
NA	0.85	0.88	0.87
p38	0.37	0.65	0.66
PARP	0.71	0.79	0.92
PDE5	0.49	0.52	0.82
PDGFRb	0.46	0.43	0.24
PNP	0.58	0.56	0.97
PPARγ	0.80	0.79	0.93
PR	0.52	0.91	0.78
RXRα	0.87	0.82	0.97
SAHH	0.92	0.92	0.94
Src	0.70	0.66	0.41
Thrombin	0.35	0.63	0.61
TK	0.86	0.81	0.94
Trypsin	0.39	0.77	0.82
VEGFR2	0.53	0.63	0.60
Average	0.62	0.73	0.75
SD	0.17	0.15	0.19

ADA, adenosine deaminase; ALR2, aldose reductase; AmpC, AmpC β-lactamase; AR; androgen receptor; CDK2, cyclin-dependent kinase 2; DHFR, dihydrofolate reductase; EGFR, EGF receptor (kinase domain); ER_agonist, oestrogen receptor (agonist-bound conformation); ER_antagonist, oestrogen receptor (antagonist-bound conformation); FGFR1, fibroblast growth factor receptor 1 (kinase domain); FXa, factor Xa; GART, glycinamide ribonucleotide transformylase; GPb, glycogen phosphorylase β; GR, glucocorticoid receptor; HIVPr, HIV protease; HIVRT, HIV reverse transcriptase; HMGR, hydroxymethylglutaryl-CoA reductase; HSP90, human heat shock protein 90; InhA, enoyl-[acyl-carrier-protein] reductase; MR, mineralocorticoid receptor; NA, neuraminidase; p38, p38 MAPK; PDGFRb, PDGF receptor β (kinase domain); PNP, purine nucleoside phosphorylase; PR, progesterone receptor; RXRα, retinoic X receptor α; SAHH, S-adenosyl-homocysteine hydrolase; SRC, tyrosine kinase Src; TK, thymidine kinase; VEGFR2, VEGF receptor 2 (kinase domain).

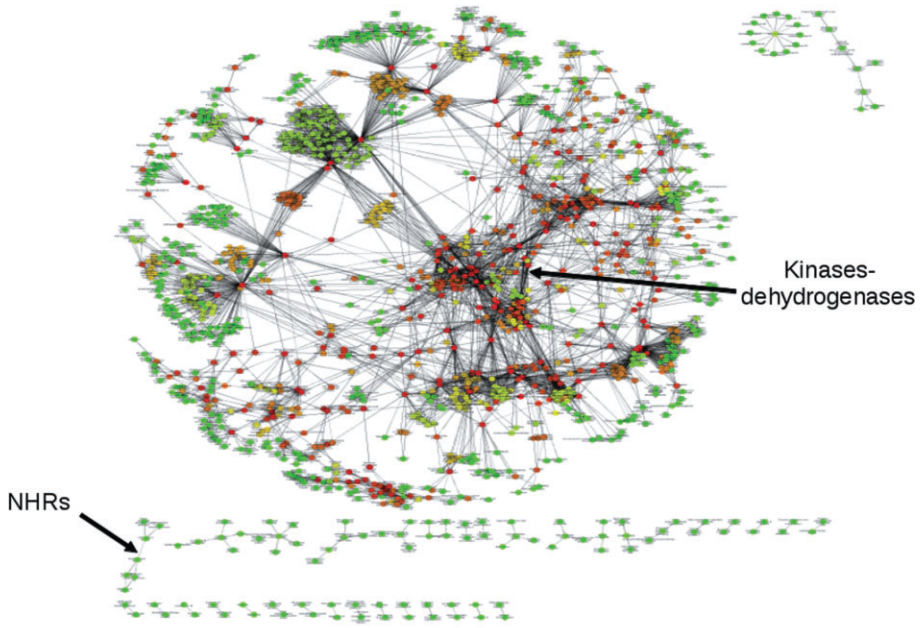


Figure 3

Small-world, broad-scale 3D receptor network with some pharmacologically relevant clusters highlighted (NHRs; kinases and dehydrogenases) and coloured by 'betweenness'.

Table 4

AUC results for the WOMBAT subset

	4D SEA	4D MAX	2D SEA
ALR2	0.56	0.67	0.42
AR	0.32	0.39	0.52
CDK2	0.51	0.67	0.69
COX2	0.58	0.81	0.43
EGFR	0.73	0.67	0.61
ER	0.65	0.73	0.66
FXa	0.50	0.43	0.77
HIVRT	0.45	0.45	0.45
p38	0.54	0.76	0.55
PDE5	0.25	0.33	0.66
PPARγ	0.83	0.81	0.72
Average	0.54	0.61	0.59
SD	0.17	0.17	0.12

ALR2, aldose reductase; AR; androgen receptor; CDK2, cyclin-dependent kinase 2; EGF receptor (kinase domain); ER, oestrogen receptor; FXa, factor Xa; HIVRT, HIV reverse transcriptase; p38, p38 MAPK.

Interestingly, the intersection of, and the difference between, the 3D target- and 4D ligand-based networks revealed that only a minimal fraction of the network is shared between them (from 0.36 to 6.80% in highly or minimally restrictive cut-off networks respectively). Low percentages may indicate that both methods yield complementary results: while the direct comparison of target ligand-binding sites could give valuable information in order to achieve some kind of target specificity, ligand-based networks could contribute with information about unexpected interactions for adverse effect prediction and polypharmacology profile optimization.

Regarding the presence of trivial information in the network, some dependence on the cut-off is observed. When the cut-off used to build the ligand-based network is low, the percentage of the network that could be considered trivial tends to increase, the score first rising due to very similar names and then, after a certain cut-off value, changing the regime to a shared-compound triviality (more than 60% of compounds in common with the other target).

In the case of 3D target networks, decreasing the cut-off has the opposite effect as it decreases the percentage of trivial edges in the network that are mostly related to a certain name similarity and independent from the number of shared compounds. This is in agreement with the origin of the network since no ligand comparisons were used to build it. In addition, the similarity in the name could be explained by the tendency of systematically naming targets that belong to the same families with similar names, which also tend to have similar functions and hence similar active (or ligand-binding) sites.

Discussion

Comparison with other methods

To our knowledge, only three other web tools are available for performing similar tasks. TarFisDock (Li *et al.*, 2006) applies the 'reverse docking' method, which consists of systemati-

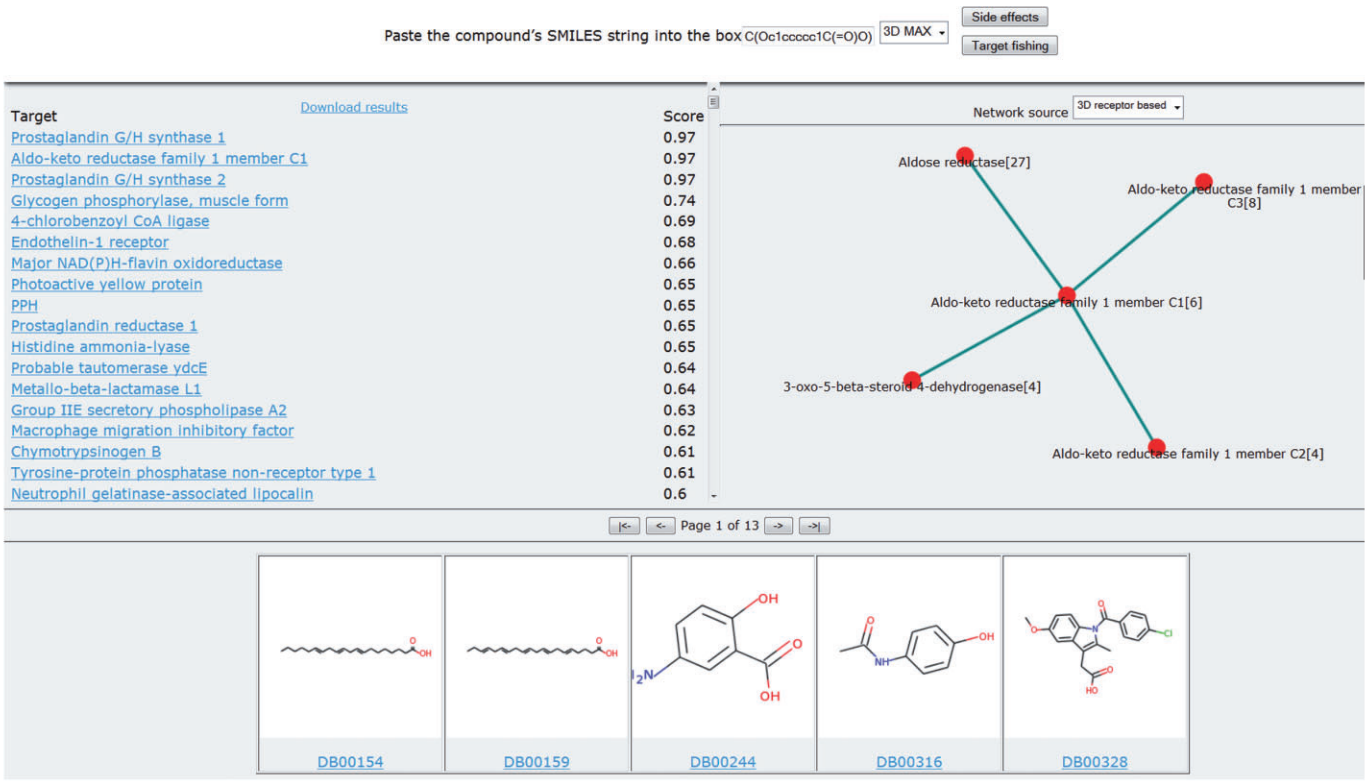


Figure 4

Screenshot of the ElectroShape Polypharmacology web server after submitting the SMILES string for acetylsalicylic acid as the query.

cally docking a ligand into hundreds of putative pharmacological target sites. This approach has the advantages of not requiring previous knowledge about true binders and not depending on functional information about the target. However, it needs a 3D model of the protein, which may be unknown or not available straightforwardly, and also suffers from the traditional problems associated with the docking methodology, namely, limited flexibility of the receptor, low speed, inaccurate scoring functions and excessive influence of the target conformation on the results. The SEA tool circumvents the docking problems by relating protein targets on the basis of the set-wise chemical similarity among all compounds that are known to interact with those targets (Keiser *et al.*, 2007). Thus, the problem is transferred to chemical space but at the expense of limiting the kind of compounds that can be predicted in a reliable manner. This method can also be used to search large chemical libraries rapidly and to build cross-target similarity maps. The third tool is STITCH 3 (Kuhn *et al.*, 2012), which allows the exploration of interactions between chemicals and targets on the basis of evidence from scientific documents. This resource currently contains interactions among 300 000 small molecules and 2.6 million proteins from 1133 organisms, and allows the visual display of interactive networks relating protein targets to which the same chemical binds.

Methodological advantages and limitations

The web tool presented here (http://ub.cbm.uam.es/chemogenomics/) implements the well-known SEA approach for target comparison using 2D ligand descriptors and also an ElectroShape comparison module that allows the estimation of 4D molecular similarities faster than other 3D methods. This is due to the pre-computation of the shape and partial charge distribution of the molecules in a format that can be stored in a database for further use, unlike the methods that need to perform n comparisons for every query. Our server also has a network explorer that allows the user to navigate the chemical (2D and 4D ligand) and the biological (target sites) regions of pharmacological space (Figure 4).

This methodology is obviously limited by the extent that pharmacological space is covered in current databases in terms of both compounds and targets. Therefore, only currently available information can be used to predict new targets or possible adverse effects for candidate molecules. It is also evident that chemical space has not been uniformly explored so that some parts of it (due to synthetic accessibility or other causes) can be better represented than others (or even not be represented at all).

Conclusions

In conclusion, we have presented here a new target-fishing approach that makes use of the ultra-fast LBVS ElectroShape methodology and is able to predict drug adverse effects, build polypharmacology profiles and relate targets from two complementary viewpoints (ligand- and target-based networks). The DUD and WOMBAT sets were employed for on-target

validation and the results were directly comparable to those obtained using other state-of-the-art target-fishing approaches. Off-target validation was performed using a limited set of non-annotated secondary targets for already known drugs. Finally, comparison of the predicted adverse effects with data contained in the SIDER 2 database showed good specificity and reasonable selectivity. All of these features are freely available from an undemanding and user-friendly web interface that (i) can be queried for both polypharmacology profiles and adverse effects, (ii) hyperlinks related targets in the three networks (2D, 4D ligand and 3D receptor) and (iii) displays the 2D structure of already annotated drugs.

Acknowledgements

A. C. C. gratefully acknowledges the FPU 2009-0203 grant and its foreign internship programme from the Spanish Ministry of Education. A. M. acknowledges financial support from Comunidad de Madrid through Fundación Severo Ochoa's AMAROUTO program. This research has been funded in part by the Spanish Comisión Interministerial de Ciencia y Tecnología (SAF2009-13914-C02-02) and Comunidad de Madrid (S2010-BMD-2457). The authors are grateful to Professor W. Graham Richards for encouragement and Dr Sree Vadlamudi for fruitful discussions during the elaboration of this work. We also would like to thank the three anonymous reviewers for their knowledgeable and insightful comments that helped to improve the quality of the original paper.

Conflict of interest

The authors declare no conflict of interest.

References

AbdulHameed MDM, Chaudhury S, Singh N, Sun H, Wallqvist A, Tawa G (2012). Exploring polypharmacology using a ROCS-based target fishing approach. J Chem Inf Model 52: 492–505.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.

Armstrong MS, Morris GM, Finn PW, Sharma R, Moretti L, Cooper RI *et al.* (2010). ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. J Comput Aided Mol Des 24: 789–801.

Ballester PJ, Richards WG (2007). Ultrafast shape recognition to search compound databases for similar molecular shapes. J Comput Chem 28: 1711–1723.

Ballester PJ, Finn PW, Richards WG (2009). Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology. J Mol Graph Model 27: 836–845.

Ballester PJ, Westwood I, Laurieri N, Sim E, Richards WG (2010). Prospective virtual screening with Ultrafast Shape Recognition: the identification of novel inhibitors of arylamine N-acetyltransferases. J R Soc Interface 7: 335–342.

Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J *et al.* (2007). Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. ChemMedChem 2: 861–873.

Berger SI, Iyengar R (2009). Network analyses in systems pharmacology. Bioinformatics 25: 2466–2472.

Berman H, Henrick K, Nakamura H, Markley JL (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35 (Suppl 1): D301–D303.

Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, Huang X-P *et al.* (2012). Automated design of ligands to polypharmacological profiles. Nature 492: 215–220.

Cabrera AC, Gil-Redondo R, Perona A, Gago F, Morreale A (2011). VSDMIP 1.5: an automated structure-and ligand-based virtual screening platform with a PyMOL graphical user interface. J Comput Aided Mol Des 25: 813–824.

Cai J, Han C, Hu T, Zhang J, Wu D, Wang F *et al.* (2006). Peptide deformylase is a potential target for anti-Helicobacter pylori drugs: reverse docking, enzymatic assay, and X-ray crystallography validation. Protein Sci 15: 2071–2081.

Ehrlich P (1907). On immunity with special reference to the relationship between distribution and action of antigens. J R Inst Public Health 15: 321–340.

Good AC, Oprea TI (2008). Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? J Comput Aided Mol Des 22: 169–178.

Grant JA, Gallardo M, Pickup B (1996). A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. J Comput Chem 17: 1653–1666.

Halgren TA (1996). Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. J Comput Chem 17: 520–552.

Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT (2010). Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. J Chem Inf Model 50: 572–584.

Hert J, Keiser MJ, Irwin JJ, Oprea TI, Shoichet BK (2008). Quantifying the relationships among drug classes. J Chem Inf Model 48: 755–765.

Hopkins AL (2008). Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4: 682–690.

Huang N, Shoichet BK, Irwin JJ (2006). Benchmarking sets for molecular docking. J Med Chem 49: 6789–6801.

Jain AN (2000). Morphological similarity: a 3D molecular similarity method correlated with protein-ligand recognition. J Comput Aided Mol Des 14: 199–213.

Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK (2007). Relating protein pharmacology by ligand chemistry. Nat Biotechnol 25: 197–206.

Knight ZA, Lin H, Shokat KM (2010). Targeting the cancer kinome through polypharmacology. Nat Rev Cancer 10: 130–137.

Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A *et al.* (2011). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res 39 (Suppl 1): D1035–D1041.

Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P (2010). A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 6: 343.

Kuhn M, Szklarczyk D, Franceschini A, von Mering C, Jensen LJ, Bork P (2012). STITCH 3: zooming in on protein-chemical interactions. Nucleic Acids Res 40 (database issue): D876–D880.

Landrum G (2011). Rdkit: Open-Source Cheminformatics. Novartis Institutes for BioMedical Research, Basel (http://www.rdkit.org/).

Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K *et al.* (2006). TarFisDock: a web server for identifying drug targets with docking approach. Nucleic Acids Res 34 (Web Server issue): W219–W224.

Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL *et al.* (2012). Large-scale prediction and testing of drug activity on side-effect targets. Nature 486: 361–367.

MDDR (2006). MDL Drug Data Report. MDL: San Leandro, CA.

Meslamani J, Rognan D, Kellenberger E (2011). sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. Bioinformatics 27: 1324–1326.

O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011). Open Babel: an open chemical toolbox. J Cheminform 3: 33.

OpenEye Scientific Software (2011). ROCS. OpenEye Scientific Software: Santa Fe, NM.

Ripphausen P, Nisius B, Peltason L, Bajorath J (2010). Quo vadis, virtual screening? A comprehensive survey of prospective applications. J Med Chem 53: 8461–8467.

Rogers DJ, Tanimoto TT (1960). A computer program for classifying plants. Science 132: 1115–1118.

Rognan D, Meslamani J (2011). Enhancing the accuracy of chemogenomic models with a three-dimensional binding site kernel. J Chem Inf Model 51: 1593–1603.

Roth BL, Sheffler DJ, Kroeze WK (2004). Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. Nat Rev Drug Discov 3: 353–359.

Rush TS 3rd, Grant JA, Mosyak L, Nicholls A (2005). A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. J Med Chem 48: 1489–1495.

Sadowski J, Gasteiger J, Klebe G (1994). Comparison of automatic three-dimensional model builders using 639 X-ray structures. J Chem Inf Comput Sci 34: 1000–1008.

Sadowski J, Schwab C, Gasteiger J (2003). CORINA, 3D Structure Generator: version 3.1, Molecular Networks GmbH, Erlangen, Germany.

Simon Z, Peragovics A, Vigh-Smeller M, Csukly G, Tombor L, Yang Z *et al.* (2012). Drug effect prediction by polypharmacology-based interaction profiling. J Chem Inf Model 52: 134–145.

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27: 431–432.

Van Der Greef J, McBurney RN (2005). Rescuing drug discovery: in vivo systems pathology and systems pharmacology. Nat Rev Drug Discov 4: 961–967.

Vidal D, Mestres J (2010). In silico receptorome screening of antipsychotic drugs. Mol. Inf. 29: 543–551.

Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH (2009). PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 37 (Suppl 2): W623–W633.

Weininger D (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28: 31–36.

Witkop B (1999). Paul Ehrlich and his magic bullets – revisited. Proc Am Philos Soc 143: 540–557.

Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B *et al.* (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34 (suppl 1): D187–D191.

Yera ER, Cleves AE, Jain AN (2011). Chemical structural novelty: on-targets and off-targets. J Med Chem 54: 6771–6785.

Zhang Y, Skolnick J (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33: 2302–2309.

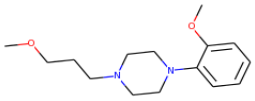
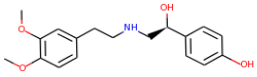
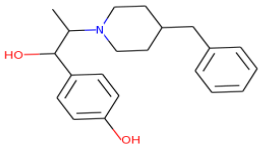
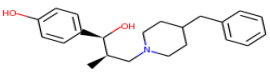
Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Table S1 Compounds used for off-target validation.
Table S2 Enrichment analysis of LBVS. BEDROC analysis ($\alpha = 20$, 80% weight for top 8%).
Appendix S1 Details on parameterization of the networks.

Supplementary Information

1. Compounds used for off target validation

Compound	SMILES	2D structure
dimetholizine	COCCCN1CCN(CC1)C2=CC=CC=C2OC	
denopamine	COC1=C(C=C(C=C1)CCNC[C@H](C2=CC=C(C=C2)O)O)OC	
ifenprodil	CC(C(C1=CC=C(C=C1)O)O)N2CCC(CC2)CC3=CC=CC=C3	
RO-25-6981	CC(CN1CCC(CC1)CC2=CC=CC=C2)C(C3=CC=C(C=C3)O)O	

2. Enrichment analysis of LBVS. BEDROC analysis (α = 20, 80% weight for top 8%)

Target	3D SEA EF1%	3D SEA BEDROC AUC	3D MAX EF1%	3D MAX BEDROC AUC	2D SEA EF1%	2D SEA BEDROC AUC
ACE	6.28	0.12	12.56	0.28	14.65	0.38
AChE	6.54	0.20	17.75	0.33	0	0.12
ADA	4.95	0.07	9.91	0.25	17.34	0.64
ALR2	3.92	0.23	39.2	0.63	3.93	0.17
AmpC	-	-	-	-	0	0.01
AR	16.64	0.34	20.48	0.63	23.04	0.46
CDK2	2.84	0.17	28.38	0.55	12.66	0.30
COMT	0	0.08	0	0.21	17.41	0.57
COX1	0	0.21	37.44	0.67	12.48	0.32
COX2	3.76	0.10	26.55	0.59	0.47	0.02
DHFR	11.43	0.48	15.32	0.48	18.72	0.94
EGFR	9.88	0.25	19.33	0.36	30.26	0.68
ER_agonist	21.19	0.50	28.76	0.64	33.30	0.48
ER_antagonist	0	0.14	17.79	0.53	0	0.48
FGFR1	0	0.00	5.80	0.11	6.62	0.10
FXa	3.42	0.06	3.42	0.06	4.11	0.13
GART	17.87	0.60	15.32	0.60	0	0.02
GPb	0	0.15	26.83	0.52	5.69	0.49
GR	3.88	0.07	9.05	0.15	14.22	0.29
HIVPr	0	0.00	6.45	0.11	3.23	0.18
HIVRT	0	0.07	0	0.07	0	0.03
HMGA	40.40	0.50	43.29	0.93	43.29	0.96
HSP90	0	0.02	27.46	0.65	13.73	0.44
InhA	5.73	0.16	26.37	0.47	0	0.14
MR	18.60	0.31	24.80	0.51	24.80	0.61
NA	4.13	0.29	14.46	0.43	19.98	0.64
P38	0.44	0.02	9.91	0.26	0	0.03
PARP	2.83	0.12	11.31	0.37	11.31	0.34
PDE5	11.18	0.21	12.30	0.30	16.77	0.60
PDGFRb	19.25	0.27	19.25	0.28	3.50	0.06

PNP	0	0.03	0	0.04	15.80	0.86
PPARγ	1.18	0.10	2.36	0.09	9.44	0.50
PR	14.38	0.23	25.17	0.63	28.77	0.35
RXRα	0	0.37	19.25	0.44	19.25	0.64
SAHH	8.95	0.51	8.95	0.56	9.09	0.78
Src	7.52	0.37	12.54	0.26	0	0.01
Thrombin	0	0.01	7.02	0.14	0	0.07
TK	9.22	0.43	27.67	0.40	23.05	0.58
Trypsin	2.06	0.05	18.51	0.50	28.80	0.51
VEGFR2	9.07	0.14	11.34	0.22	0	0.09
Average	-	0.20	-	0.39	-	0.37
st. deviation		0.16		0.21		0.28

3. Details on parameterization of the networks.

SEA parameters based on the DrugBank database were derived following the method described by Hert et al. [1]; the power-law equation was fitted used the curve fitting module of OpenOffice.org.

Regarding 3D receptor network, binding pocket definitions were adopted from the sc-PDB database where they are available as separate files.

Triviality name measurements were calculated using the *SequenceMatcher* method included in Python's *difflib* module.

[1] Hert J, Keiser MJ, Irwin JJ, Oprea TI, Shoichet BK (2008). Quantifying the relationships among drug classes. *J. Chem. Inf. Model.* **48**(4)**:** 755-765.

Article V

AtlasCBS: a web server to map and explore chemico-biological space

Background and author's contribution

Pharmaceutical chemistry lies at the interface between chemistry and biology. Practitioners should explore chemical space looking for feasible molecules that could present favorable physicochemical properties while ensuring that a reasonable biological activity is achieved. This daunting task leads to some kind of exploration of the resulting chemico-biological space (CBS). Usually, this process is propelled by a systematic sampling of the chemical space accessible to synthetic chemists and corrected rounds after rounds by the binding information obtained after evaluation of the new compounds. Project analyses are difficult due to the lack of a systematic framework that can evaluate simultaneously the chemistry and the biology.

An increasing number of databases have been made available (e.g. ChEMBL, PubChem, BindingDB, MOAD, etc.) which contain ligand information and their target-binding profiles. These databases contain nothing but a large description of several CBS available in the literature. However, this information could not be easily represented and tends to be shown as data tables.

In this article we present a new web tool to perform generic analyses of present and past projects. Using very simple efficiency variables (SEI, BEI, nBEI, mBEI, NSEI, NBEI and LE) the tool allows to map the CBS in planes of efficiency that simplify the project information and show a clear pathway to the most promising candidates avoiding possible pitfalls in the process. These planes hold the physicochemical properties of the compounds (polarity, weight) and the binding information for a given target (K_i, K_d, IC_{50}). Retrospective analyses showed that suitable molecules tend to be located in the upper-right corner of the planes where good efficiency relative to the weight and also good efficiency relative to the polarity could be found. The tool also allows the analysis of general databases such as PDBBind, ChEMBL and BindingDB in an atlas-like manner and the comparison with personal project information and compounds.

The main author of the manuscript designed the interface and developed the server. The manuscript was written collectively by all the authors.

J Comput Aided Mol Des (2012) 26:995–1003
DOI 10.1007/s10822-012-9587-5

AtlasCBS: a web server to map and explore chemico-biological space

Álvaro Cortés-Cabrera · Antonio Morreale · Federico Gago · Celerino Abad-Zapatero

Received: 26 March 2012 / Accepted: 30 June 2012 / Published online: 14 July 2012

Abstract New approaches are needed that can help decrease the unsustainable failure in small-molecule drug discovery. Ligand Efficiency Indices (LEI) are making a great impact on early-stage compound selection and prioritization. Given a target-ligand database with chemical structures and associated biological affinities/activities for a target, the AtlasCBS server generates two-dimensional, dynamical representations of its contents in terms of LEI. These variables allow an effective decoupling of the chemical (angular) and biological (radial) components. BindingDB, PDBBind and ChEMBL databases are currently implemented. Proprietary datasets can also be uploaded and compared. The utility of this atlas-like representation in the future of drug design is highlighted with some examples. The web server can be accessed at http://ub.cbm.uam.es/atlascbs and https://www.ebi.ac.uk/chembl/atlascbs.

Á. Cortés-Cabrera · F. Gago · C. Abad-Zapatero
Departamento de Farmacología, Universidad de Alcalá, 28871 Alcalá de Henares, Madrid, Spain

Á. Cortés-Cabrera · F. Gago · C. Abad-Zapatero
Unidad Asociada Instituto de Química Médica (CSIC), Madrid, Spain

Á. Cortés-Cabrera · A. Morreale
Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Campus UAM, c/Nicolás Cabrera 1, 28049 Madrid, Spain

Present Address:
C. Abad-Zapatero (✉)
Center for Pharmaceutical Biotechnology, University of Illinois at Chicago, MBRB building (MC870), Chicago, IL 60607, USA
e-mail: caz@uic.edu

Keywords Ligand efficiency indices · Chemico-biological space · Structure–activity databases · AtlasCBS server · Efficiency-based drug design · Efficiency planes

Introduction

The ever growing advances in the fields of structural biology, high-throughput screening and structure-based drug design have resulted in an exponential increase of the information related to targets, ligands, and their complexes that is stored in several databases (i.e., Structure–Activity-Relationship or SAR databases: BindingDB [1], ChEMBL [2], PDBBind [3], among others). The vastness of chemical space as it relates to medicinal applications has been recognized [4] and certain tools to aid in navigating it have been introduced [5, 6]. Nonetheless, an effective method to combine biological targets with the subset of ligands with which they interact and map them in chemico-biological-space (CBS) is lacking.

In the last few years the concept of ligand efficiency has taken hold in medicinal chemistry following the pioneering work of Hopkins et al. [7] who suggested an efficiency-based reference for lead selection taking into account the size of the ligand. The concept was extended to include efficiency in two complementary variables: size (molecular weight, MW, of the ligand) and polarity (Polar Surface Area, PSA) [8]. In this formulation, the size variable was defined as BEI = pK_i/(MW/1,000) and the polarity variable as SEI = pK_i/(PSA/100) where K_i is the binding affinity of the inhibitor (see Table 1). BEI is the binding efficiency index relating potency to MW on a per kDa scale and SEI is the surface efficiency index monitoring the potency gains as related to the increase in PSA referred to 100 $Å^2$. It was then proposed that ligand efficiency indices

Table 1 Definitions of ligand efficiency indices used in AtlasCBS

Acronym	Definition
BEI	pK_i/MW, pK_d/MW, pIC_{50}/MW[a]
SEI	pK_i/PSA, pK_d/PSA, pIC_{50}/PSA[b]
NSEI	pK_i/NPOL[c]
NBEI	pK_i/NHA[d]
nBEI	$-\log_{10}$ [K_i/NHA]
mBEI	$-\log_{10}$ [K_i/MW]

Examples of possible efficiency planes represented and available in the AtlasCBS server and a brief description of their characteristics and appearance

1. SEI, BEI (x, y). Slope of the lines 10*(PSA/MW). No intersect

2. NSEI, NBEI (x, y). Slope of the lines NPOL/NHA: a rational number. No intersect

3. NSEI, nBEI (x, y). Slope of the lines NPOL, intersect $\log_{10}$(NHA)

4. NSEI, mBEI (x, y). Slope of the lines NPOL, intersect $\log_{10}$(MW)

Any combination is possible but pairs of variables related to affinity/polarity (SEI-like, x) and affinity/size (BEI-like, y) are strongly recommended. With this choice of variables the polarity or physico-chemical characteristics (as given by PSA/MW or NPOL/NHA) of the chemical compounds increases counterclockwise as indicated above, mimicking a graphical representation of the variables considered in Lipinski's Rule of Five (Ro5) [13, 17] by changes in the angular coordinate (slope) of the lines

[a] MW: molecular weight (in kDa). $pK_i = -\log_{10} K_i$

[b] PSA: polar surface area scaled to 100 $Å^2$

[c] NPOL: number of polar atoms (N and O)

[d] NHA: number of heavy atoms (non-hydrogen)

(LEI) could be used to guide drug discovery. These initial ideas have been extended to include other indices such as those related to lipophilicity of the ligands: LLE, lipophilicity ligand efficiency; LLE_{Astex}, lipophilicity ligand efficiency for fragments; and LELP, lipophilicity in optimization and drug-likeness. In addition, the ligand efficiency decreases with the size of the ligand and thus size-corrected (or size-independent) ligand efficiencies (SILE) have also been introduced. A comprehensive review of the various LEI mentioned has been published recently [9], where the original concepts and the newly introduced variables have been fully discussed and documented in the context of the current practice of medicinal chemistry .

The concept of an atlas-like representation of CBS based on the use of two complementary LEI was introduced recently [10]. The proposed formulation allowed a representation in Cartesian planes (efficiency planes) that combined the affinity of the ligand towards the target and two important physico-chemical properties of the ligand: size and polarity. Briefly, a combination of efficiency indices defined in size (BEI, y) and polarity (SEI, x) can be represented on two-dimensional (2D) efficiency planes (see Table 1 for a summary of proposed variables). The rationale was that, as a first approximation, the drug discovery process attempts to optimize the combination of three variables: potency, size and polarity and therefore a graphical representation of the combination of these parameters could aid in the optimization process. Given a choice of axes and definition of variables, the appearance of the plots is characterized by a counterclockwise increase of the polarity of the compounds in the angular coordinate (chemistry) and a radial distance from the origin related to the affinity towards the target (biological) (see Table 1). Applications of this representation to several areas of drug discovery have also been described [10].

However, thus far, no friendly tool is available to connect 2D or 3D ligand structures plus their chemical properties (chemical space) with biological affinity/activity data pertaining to one or more target proteins (biological space). Moreover, a friendly tool is lacking that would allow the medicinal chemistry community to explore the value of the graphical representation of CBS and the efficiency (in size and polarity) of their chemical entities towards a certain target. The AtlasCBS server introduced and described here is such an application.

Materials and methods

LEI and molecular properties calculation

Several LEI are calculated, as described elsewhere, using the formulas contained in Table 1 and in Ref. [10]. Molecular properties such as atomic masses, number of polar (NPOL), non-hydrogen atoms (NHA), and PSA are polar surface area are calculated using the Chemistry Development Toolkit (CDK) [11, 12].

Web design and usage

The main web page for the server contains the five tabs shown in Fig. 1a, b: (a) **Main**: basic information about the server and its purpose, main references, contact information, and access to the main topics covered in the **Help** tab; (b) **Map viewer**: tools for uploading the data from existing databases and for visualizing and analyzing their content; (c) **Login**: required only if the user wants to have private database access; (d) **Help**: information on how to use the AtlasCBS server; and (e) **About**, details concerning the institutions and people involved in the project. The **Login** tab changes to **Manage data** upon a successful login allowing access to proprietary user data. Underneath the server there are three modules that provide all the functionalities for the tabs:

- **(A) Map viewer**. The graphical engine of the server represents data from different sources, allows visualization

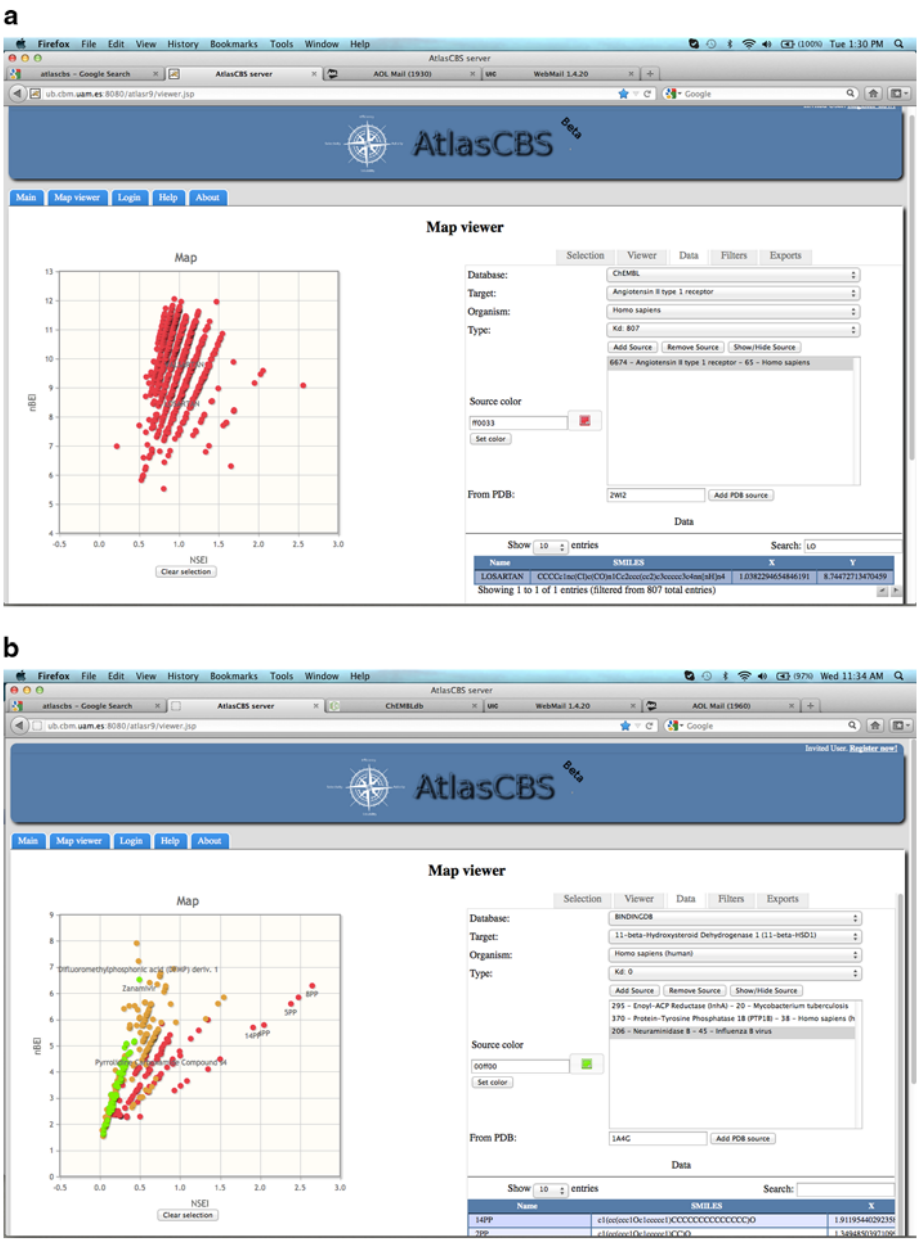


Fig. 1 Example screens presented by the AtlasCBS server. **a** The *upper left* tabs correspond to the main pages of the server: **Main**, **Map Viewer**, **Login** (for private access), **Help**, and **About** (see text). The left graphical panel within the page represents a typical efficiency plane (NSEI vs. nBEI) for the 807 entries found in ChEMBL for Angiotensin II receptor type I with K_d affinity values (see Type window). Each point in the plane represents a target-ligand pair. The angular coordinate (NPOL in this case) corresponds to the number of polar (N,O) atoms of the ligand increasing counterclockwise (NPOL = 3 12). The radial coordinate corresponds to the affinity of the ligands towards the target. The *top right* panel shows the different options for the management of the session within the **Map viewer** and choice of database, target and organism. Most importantly, the choice of LEI as Cartesian axes (x, y) (within **Viewer**) determines the appearance of the efficiency planes. The *lower right* panel shows the compounds selected in an alphabetical list. The values of the LEI variables are also shown in this window along with the SMILES strings of the corresponding compounds. The SMILES string of one of the selected compounds (LOSARTAN) is shown at the bottom, resulting from inputting 'LO' in the search box. It is also annotated on the left map in black fonts. Using the 'Duplicate window' tab, multiple windows can be displayed simultaneously to compare pages in the 'AtlasCBS' with different variables or scales, as in a real life atlas. **b** As in **a** but in this case the map in the NSEI-nBEI plane corresponds to the response of the server to three succesive PDB access codes: 2H7L, enoyl-ACP-reductase (InhA) from *Mycobacterium tuberculosis* (*purple*), 2CNE, human protein tyrosine phosphatase 1B (*amber*), and 1A4G, neuraminidase B (*green*). The colors were changed for contrast using the 'Source color' option. In addition to the compound markers corresponding to the three PDB entries, other compounds have been *highlighted* (black lettering 14PP, 7PP etc.). Of note, the marketed drug zanamivir occupies the top (most efficient) position of the neuraminidase B inhibitors (*green set*) (see Fig. 3 for a close-up)

of chemical structures, and provides efficient filtering and searching tools to compare and classify molecules and their efficiencies. The application opens in the **Data** tab within the **Map viewer** panel. First, the users should select any data source (target and organism) available within the AtlasCBS server, extracted from a previous release of BindingDB (www.binding.org), PDBBind (www.pdbbind.org) and ChEMBL (https://www.ebi.ac.uk/chembl), scrolling the corresponding server tabs, or load an external data set (see below). Any external set should include the compound name, structural description of the molecules (SMILES strings) and their affinity/activity (K_d, K_i, or IC_{50}) values.

1. **Data selection.** The user selects from the available data (K_i, IC_{50} or K_d) in the databases and uploads (**Add Source** tab) the data to the viewer. By default, the first map shown is nBEI versus NSEI, where the appearance is a set of lines of slope given by the number of polar (N + O) atoms, increasing counterclockwise. To generate other maps, the user selects x and y variables in the **Viewer** tab from the sets: SEI, BEI; NSEI, NBEI; NSEI with nBEI or mBEI, respectively. A brief definition of the different variables in tabular form is available by opening an adjacent pull-down tab, and other details are provided in a link. Any combination is possible but complementary pairs (LEI per size and polarity: y and x, respectively) are recommended. Given a map, molecules can then be selected by clicking on them on the map or from the list. The 2D structure of the selected molecule can be seen within the **Selection** tab together with its basic physico-chemical properties. A list of compound names, SMILES strings and LEI values (x, y) for the displayed compounds is shown on the lower right panel (Fig. 1). Molecules can be selected by a simple character search on the same panel or filtered (within the **Filter** tab) using either a "range of values" or the "Slope" option to choose those that share the same number of polar atoms (in the NSEI-nBEI plane). Selected compounds can be compared or used for similarity searches employing molecular fingerprints and Tanimoto coefficients. Other features include: (a) mixing and visualization of different data sources at the same time; (b) changing colors for the different data sources, with 'on' and 'off' capabilities; and (c) dynamic scaling of the axis to zoom in/out on particular areas of interest (Fig. 1). These features are intended to give the sense and the feel of a geographical atlas. Finally, it is possible to save and restore any working session as well as to export 'working datasets' and plots of specific efficiency planes of interest. The appearance of some of these options could be dependent on the browser's capabilities (Table 2; Figs. 1, 2).
2. **Relation to other databases.** Within **Data,** it is possible to input directly a PDB access code and the server will make available to the user all the affinity data present in BindingDB for that target. The activity or affinity parameter (IC_{50}, K_i, or K_d) with the larger number of entries is selected by default. As before, once a compound has been selected and appears in the selected list, the **Selection** tab, within the **Map Viewer** module, shows the compound's 2D structure together with its name, basic physico-chemical properties (MW, number of polar atoms, PSA, number of heavy [i.e. non-hydrogen] atoms) and a direct link to BindingDB (Fig. 2).

- **(B) Manage data (Private database manager).** This feature allows users to upload and process their own datasets in a secure way provided they register and accept the terms of usage of the site (access is granted using a valid e-mail address and a user-chosen password).
  1. Datasets containing affinity data are read in as semicolon-separated values ("CSV"), which are readily available from common spreadsheets such as those produced by Microsoft Excel or OpenOffice Calc. Compounds can also be added manually as long as the required data are correctly given in the specified format. Users can modify each field interactively and they can also use filtering and searching tools.
  2. Chemical library contents can be uploaded into the private data stream and the server will generate randomized affinity (K_i) values to explore/simulate a screening experiment (see below).

Table 2 Viewers characteristics

Characteristic	Java applet viewer	Javascript/HTML viewer
SMART filtering	✔	–
Similarity search	✔	–
Simple filters	✔	✔
Polarity highlighting	✔	–
Labels	✔	✔
Mixed sources	✔	✔
Save session	–	✔
Dynamic scale change	✔	✔

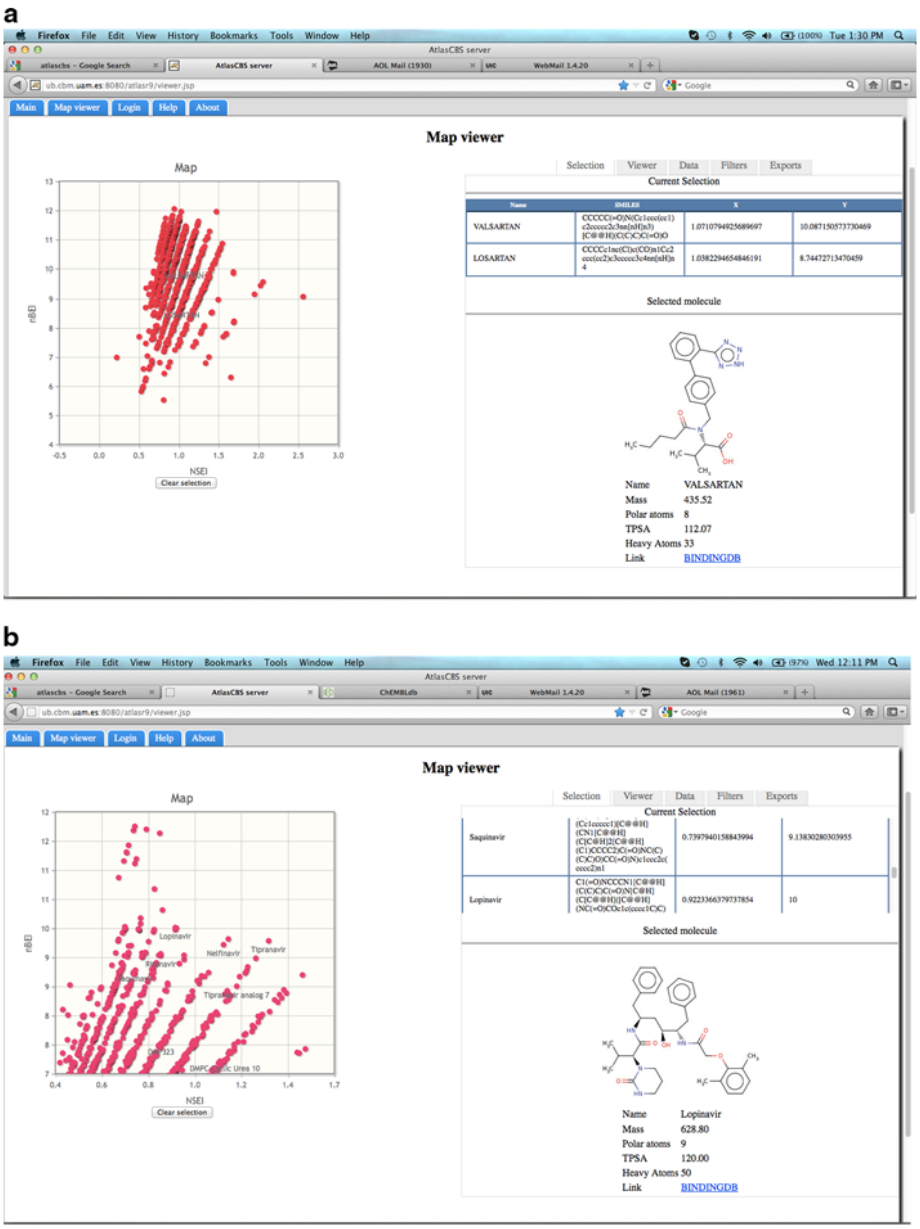


Fig. 2 Mapping efficiency and structure in chemico-biological space. **a** Image from the **Map Viewer** panel illustrating the position of a certain selected compound (VALSARTAN) targeting angiotensin II type 1 receptor (as in Fig. 1a) and the corresponding 2D structure within the **Selection** tab. The basic physico-chemical parameters of the compound are displayed as well as a direct link to BindingDB that can be used to look up some additional information. The relative efficiency of other compounds for the same target such as LOSARTAN (Fig. 1a) can be compared. **b** Close-up image from the **Map Viewer** panel showing a subset of the compounds extracted from BindingDB as a response to the PDB access code 1OHR (HIV-1 protease in complex with nelfinavir). The 2D structure of the selected compound is shown. The progressive migration towards higher efficiencies of the compounds targeting HIV-1 protease can be seen. The structure of Lopinavir, a later HIV-1 protease inhibitor to make it into the market, is shown for reference. The angular component (*slope of the lines*) relates to the number of polar atoms (N + O). The radial coordinate is given by the measured affinity towards the corresponding target. Different measurements for different targets (i.e. wild-type vs. mutant(s)) will all map along the same line. The thickness of the lines depends on the number of compounds with the same NPOL value but different number of non-hydrogen atoms ($\log_{10}$ (NHA)). See Table 1. The selection of this area (NSEI: 0.4 1.7; nBEI: 7 12) was made by highlighting a certain region of the full NSEI-nBEI plane with the mouse following representation on the plane of all the compounds with affinity for the HIV-1 protease target. The physico-chemical properties of the ligands are shown together with a direct link to BindingDB as above. Note how lopinavir, tipranavir and nelfinavir occupy the highest efficiency positions for the corresponding NPOL lines in relation to other analogs (e.g. tipranavir analog 7, *third line* from the *right*)

- **(C) Help.** It contains and explains the elements and functionalities of the server as well as references to background papers. There are also references to previous papers in the main portal of the server.

Implementation details

The server is organized in three layers: clients, application server and database. Each layer can communicate with the nearest neighbor but not beyond. The three layers have been implemented with the following elements for the different components, respectively: (1) Java, JavaScript, and HTML clients; (2) the Apache Tomcat servlet container; and (3) the MySQL database engine. The front-end is based on HTML, JavaScript and Java Server Pages (JSP) or a Java applet; Java servlets handle the data traffic between the interface and the database, using a Model-View-Controller (MVC) paradigm. The information contained in a current release of BindingDB (19/05/2012), PDBBind (v2011) and ChEMBLdb (v13) was imported into the MySQL server's database by standalone Java programs that also compute the molecular properties and the efficiency indices for the molecules. Data for three affinity parameters (K_i, IC_{50}, K_d) are extracted and retained within the server connected to the appropriate target and organism. In the case of BindingDB, the SDF (Structure Data File) was converted to SMILES format using RDKit (http://www.rdkit.org/) and then inserted into the database with the information provided for targets and organism in the property fields of the same SDF file, along with the calculated molecular properties by CDK [11, 12]. For PDBBind the data were transformed directly into SMILES strings from the PDB files and the affinity information in the annotated set was used. Finally, for ChEMBL, a subset of the original release for MySQL was imported, excluding extreme affinity/activity values and considering only data in nanomolar units for K_i, IC_{50} and K_d. Names or reference IDs for molecule in those databases were recorded in order to provide links from the platform to the original data sources.

Users-uploaded external data are processed on-the-fly, with a special servlet using the CDK and the AJAX technology [12]. The accepted format of the user's external database and examples are described in the help pages. For each entry, the CSV-formatted file should contain molecule name, SMILES code, type of affinity/activity variable (K_i, IC_{50}, K_d), and the affinity/activity value (in nanomolar concentration units).

A separate tab is available to upload the contents of proprietary chemical libraries (without any affinity data) within the **Manage data** tab. The files containing the chemical composition of the libraries should be in SMILES format separated by a tab character from the compound name. Only two items are required: compound number (e.g. AS0045) and the corresponding SMILES string. The server assigns internally random affinity (K_i) values to the compounds in the micro- to nanomolar range. The physicochemical properties of the compounds within the library are represented by the range of the angular coordinates in the efficiency planes. The regions of the efficiency plane accessible to different affinities are simulated by the random values of the affinity constants (radial direction) [13].

Visualization of the data is based on a Java applet that enables graphical representation and allows the display of multiple pages simultaneously, zooming in and out of user's predefined areas, and selection of compounds using SMARTS strings, similarity, or automatic detection. In the NSEI-nBEI (x, y) plane, compound series are easily followed by the slope of the lines that correspond to the number of polar atoms (N + O), which increases counterclockwise. We have also developed a javascript-based web application that avoids the use of Java but implements fewer features, although the result is browser-dependent. For a list of the features supported by each visualization module see Table 2.

Molecular fingerprints and Tanimoto coefficients

Molecular comparisons are based on the calculation of the CDK [11, 12] fingerprint, which encodes the topology of the compounds as bit strings. Bit strings are compared using the Tanimoto coefficient (Tc, Eq. 1), which evaluates, from 0 (not at all) to 1 (identity), the similarity between the compounds:

$$Tc = Nab/(Na + Nb - Nab) \quad (1)$$

where Nab are the bits in common for two compounds (a, b) and Na and Nb the number of bits activated in compounds a and b, respectively.

Results and discussion

LEI represent a relatively simple concept that naturally connects the chemistry and the biology of receptor-binding ligands via one or more affinity/activity variables (K_i, IC_{50}, or equivalent). Other expanded definitions of ligand efficiency proposed recently to expand the initial ideas [9] could further impact the future of drug discovery. In this work, we have developed an application based on a unified formulation of LEI (see Table 1 for definitions) that are calculated by weighting the affinity values with molecular properties such as the number of heavy (non-hydrogen, NHA) or polar atoms (NPOL = Number of N + O), the polar surface area (PSA), or the molecular weight (MW).

The combined use of two complementary LEI, namely (1) affinity/polarity (K_i combined with NPOL, PSA), x-axis, and (2) affinity/size (K_i combined with NHA, MW), y-axis, allows a very intuitive depiction of the database contents as a series of Cartesian diagrams that constitute an atlas-like representation of CBS. The characteristics and appearance of these plots ("efficiency planes") depend on the choice of variables (see Table 1) and can be examined using the AtlasCBS server.

A very interesting characteristic of these LEI variables is that pairs of complementary indices can be displayed on an efficiency plane that unveils some of the intricacies of CBS in an appealing graphical manner. Although any combination of affinity/polarity (x-coordinate) and affinity/size (y-coordinate) could be useful [10], of special interest are the NSEI-nBEI, NSEI-mBEI (x, y) efficiency planes (Table 1; Figs. 1, 2, 3). In this type of plot, the slope of the line occupied by each target-ligand pair (its angular coordinate) depends only on the chemical composition of the ligand (in this case given by the number of polar atoms). The unique position along the line (the radial coordinate) for each target-ligand pair will depend on the biological affinity/activity of the ligand for/on the given target [10]. The drug discovery efforts can then be represented on the suggested efficiency planes as 'trajectories' in CBS [10] and optimized trajectories could be devised or proposed in the future based on numerical or statistical criteria [14].

It has been shown in initial retrospective studies that the optimized ligand(s) within a series of analogues typically map on the upper, right-hand, quadrant of the efficiency planes (e.g. NSEI and nBEI, or similar), where both variables are maximized [10, 13, 14]. Examples are shown in Figs. 1b, 2b and 3. Therefore, in the near future, it is theoretically possible to envisage an automatic procedure that will optimize ligand efficiencies by replacing some molecular fragments and evaluating the LEI of the new candidates in an iterative fashion, until the best possible ligand(s) is(are) found [15]. It is precisely here that the future power of this methodology as a graphical and numerical guide in the search for better drug candidates should be apparent. In the future, other definitions [9], and possibly more variables, could be incorporated into this graphical framework to strengthen the prospective value of these concepts.

Other combinations of the variables defined in Table 1 can be selected and used in pairs to build up an "electronic atlas" composed of different "Cartesian maps" (i.e., pages or efficiency planes). These maps will depict complete views or selected regions of CBS at different scales, constituting what we refer to as an atlas-like representation of CBS (Fig. 2b). This AtlasCBS server is visual and dynamic

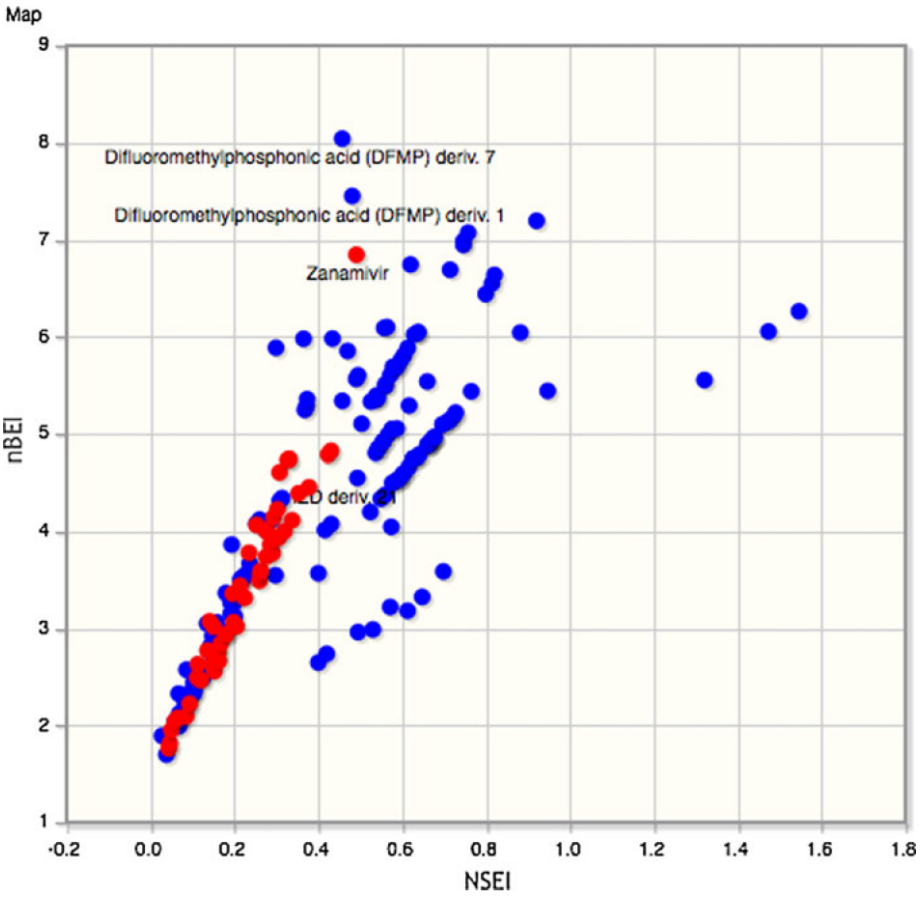


Fig. 3 Efficiency plane representation of the compounds with affinity for neuraminidase and PTP1B. The information was extracted from the BindingDB database upon entering the PDB access codes 1A4G (*red*) and 2CNE (*blue*) in the AtlasCBS server application. The plot illustrates how the marketed compound (zanamivir) optimizes the efficiency of the ligand in size (y-axis) and polarity (x-axis). This image was exported using the '**Exports**' option from the AtlasCBS server after removing the Enoyl ACP-reductase data shown in Fig. 1b

by nature and we would expect it to be extremely useful to help navigation through the 'vastness of chemical space' [4]. To the best of our knowledge, this is the first time a tool is presented that graphically displays, and naturally maps and classifies in a user-friendly way, the information stored in these ligand-receptor databases in terms of LEI. Apart from representing the chemico-biological content of public databases such as BindingDB, PDBBind and ChEMBL, our server also allows interested users to upload, map and compare their own proprietary datasets. Thus, the CBS of known targets can be easily explored and data from different sources can be visualized and compared. The URL for the server is http://ub.cbm.uam.es/atlascbs but a mirror is also available at the European Bioinformatics Institute (EBI) Hinxton campus server https://www.ebi.ac.uk/chembl/atlascbs. Both sites require registration only to grant secure and confidential access.

Some applications of the AtlasCBS concept have been reviewed lately by Abad-Zapatero and Blasi [10, 13] in several domains of the drug discovery process. Examples are the analysis and comparison of the contents of different databases (mapping of drugs vs. non-drugs), polypharmacology, fragment-based ligand design strategies, drug discovery trajectories and others. The AtlasCBS server presented here allows the exploration of these important areas of drug discovery dynamically on-line for the first time. We wish to encourage the drug discovery community to use this tool so that it can be further improved.

Besides the contents of the three SAR databases currently included in the server (BindingDB, PDBBind and ChEMBL), the internal structure of the application allows a direct link with the Protein Data Bank (PDB) and BindingDB in two effective ways. From the **Data** tab, the user can input a PDB accession code containing a target-ligand complex (for example 1A4G, for the target neuraminidase in complex with zanamivir). The application will use the information in the PDB entry to extract all the corresponding affinity data for that target from BindingDB and will represent the available ligands on the nBEI versus NSEI efficiency plane, with the entered compound highlighted for reference (see Figs. 1b, 2b, 3). A direct link with the information for any compound present in BindingDB is possible from the **Selection** tab and this facilitates the access to further information about the target-ligand pairs under study.

The examples presented above of the efficiency data for compounds targeted towards HIV-1 protease (Fig. 2b, extracted using PDB code 1OHR) and influenza virus neuraminidase (PDB code 1A4G, Fig. 3) illustrate a very important use of the efficiency planes presented by the AtlasCBS tool that could have an impact on future drug discovery efforts. Namely, that compounds with maximal efficiencies in size and polarity are often the best suited preclinical candidates for further development and often correspond to the successful marketed drug, as supported by other retrospective studies [13–15]. This notion can be extensively explored using the server as the available data can be represented in a variety of efficiency planes. To make it more effective and complete, the inclusion of additional LEI into the AtlasCBS framework is being considered and will be implemented in the near future. This will include size-related indices such as the original and commonly used ligand efficiency definition proposed by Hopkins et al. [7], and also size-independent (e.g. SILE) and polarity-related indices (e.g. LLE and others [9]).

Undoubtedly, the most interesting applications will be those for which LEI are used prospectively to guide the drug discovery process. Our suggestion would be to incorporate routinely the LEI framework into the drug discovery pipeline. In a recent study, Blasi et al. [16] devised a workflow to obtain better drug candidates targeting the transthyretin carrier protein (TTR) by combining LEI, pharmacophoric search and ligand docking. Briefly, a retrospective NSEI-nBEI map was first built with some known binders so as to select the most appropriate candidate for further improvement. Second, the core structure of the selected compound was used as a pharmacophore to search into a database of commercially available compounds. Those molecules fulfilling the pharmacophoric requirements were submitted to docking and the 80 top-scoring hits were selected. Third, these scores were transformed into estimated K_i values for calculating their 'theoretical' LEI. Finally, a prospective map was built and the 12 'most efficient' compounds (those having the highest values of NSEI-nBEI for the different NPOL lines) were selected for experimental tests of activity and pharmacokinetic behavior. The results with the compounds proposed for NPOL = 5 (four compounds, the most polar and easier to synthesize) so far confirm the prediction of being the most efficient in the experimental assay (Blasi and Quintana, personal communication). Using as a guide the suggestion that compounds with maximal efficiencies are likely to be good candidates for further development, we propose that the above strategy and the use of the AtlasCBS server would be advantageous to the drug-discovery community. This could set the basis for a more rigorous, numerically and efficiency-based drug discovery paradigm [15].

Conclusion

An effective web tool is presented that aims to facilitate the drug discovery process by providing an atlas-like representation of the CBS using LEI as descriptors. This web server allows the graphical visualization of database

contents as pages in a map-like environment, with different variables and scales. The CBS can be easily navigated to examine the efficiency of existing and prospective target-binding molecules differing in size and polarity. We propose that the atlas representation can be extremely useful as a guide in several areas of drug discovery, including mapping design efforts, exploring new design strategies and optimizing candidates in hit-to-lead campaigns.

Acknowledgments The assistance of the ChEMBL group in facilitating the data extraction for the AtlasCBS database is greatly appreciated. This work was supported by grants from Comunidad Autónoma de Madrid (S-BIO-0214-2006 and S2010-BMD-2457 to F.G. and A.M.), Fundación Severo Ochoa (AMAROUTO program to A.M.) and the Spanish Ministerio de Educación (FPU AP2009-0203 to A. C. and SAB2010-0037 to C. A-Z). The suggestions and comments from the reviewers of this work are greatly appreciated. The insightful comments, discussions and suggestions of Daniel Blasi of the Platform of Drug Discovery (Dr. J. Quintana, Director) within the Parc Científic Barcelona are also appreciated. We gratefully acknowledge the help of Mark Davies, John Overington and Peter Rose in our efforts to relate AtlasCBS to ChEMBL and PDB.

References

1. Liu T, Lin Y, Wen X, Jorissen RN, Gilson M (2007) Nucleic Acids Res 35(Database issue):D198
2. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2011) Nucleic Acids Res 40(Database issue):D1100
3. Wang R, Fang X, Lu Y, Yang CY, Wang S (2005) J Med Chem 48(12):4111
4. Lipinski C, Hopkins A (2004) Nature 432(7019):855
5. Oprea TI, Gottfries J (2001) J Comb Chem 3(2):157
6. Watson P, Verdonk M, Hartshorn MJ (2003) J Mol Graph Model 22(1):71
7. Hopkins AL, Groom CR, Alex A (2004) Drug Discov Today 9:430
8. Abad-Zapatero C, Metz JM (2005) Drug Discov Today 10(7):464
9. Hann MM, Keseru GM (2012) Nat Rev Drug Discov 11(5):355
10. Abad-Zapatero C, Perisic O, Wass J, Bento AP, Overington J, Al-Lazikani B, Johnson ME (2010) Drug Discov Today 15(19–20):804
11. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) J Chem Inf Comput Sci 43(2):493
12. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL (2006) Curr Pharm Des 12(17):2111
13. Abad-Zapatero C, Blasi D (2011) Mol Inform 30(2–3):122
14. Christmann-Franck S, Cravo D, Abad-Zapatero C (2011) Mol Inform 30(2–3):137
15. Abad-Zapatero C (2007) Expert Opin Drug Discov 2(4):469
16. Blasi D, Arsequel G, Valencia G, Nieto J, Planas A, Pinto M, Centeno NB, Abad-Zapatero C, Quintana J (2011) Mol Inf 30(2–3):161
17. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Adv Drug Deliv Rev 23:3

Article VI

A computational fragment-based de novo protocol based on pseudo-Murcko fragmentation and ligand efficiency indices

Background and author's contribution

Computational fragment-based tools are available since the early 90s. LUDI [1] was a pioneer in the field, implementing a tool that is still in use today and gave birth to many other programs based on its philosophy and methods.

Protocols of this kind suffer from three main problems: (i) the proper prediction of binding energies from the putative compounds, (ii) the resulting physichochemical profiles of the compounds, and (iii) the synthetic accessibility of the output molecules. The first one can be addressed with improved scoring schemes built specifically for the detail necessary for the fragments. The second one was not addressed until recently and could be improved using certain *tricks* such as rule-of-five [2] filters or ligand efficiency indices. The third and last depends on the nature of the fragment database itself and the rules used to connect new fragments to an existing molecule. It implies using fragmentation rules that take into account possible chemical reactions to connect them, such as in RECAP [3].

In this chapter we describe the implementation of a growing fragment protocol with a modified fragmentation scheme and an approach driven with ligand efficiency in mind.

The main author wrote the applications and scripts, and the manuscript draft in collaboration with the other authors.

A computational fragment-based *de novo* design protocol based on pseudo-Murcko fragmentation and guided by Ligand Efficiency Indices (LEIs)

Álvaro Cortés-Cabrera[1,2], Federico Gago[2] and Antonio Morreale[1,3,*]

[1]Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Campus de Cantoblanco UAM, E-28049 Madrid, Spain.
[2]Área de Farmacología, Departamento de Ciencias Biomédicas, Unidad Asociada de I+D+I al CSIC, Universidad de Alcalá, Alcalá de Henares, E-28871 Madrid, Spain.
[3]Current address: Repsol Technology Center, Móstoles, E-28923 Madrid, Spain.

Keywords

Fragment-based drug design, ligand efficiency indexes,

Abstract

[*]**Corresponding author:** Antonio Morreale. Repsol Technology Center, Ctra. de Extremadura, A-5, km 18, E-28935 Móstoles, Madrid, Spain. E-mail: antonio.morreale@repsol.com. Telephone number: + 34 917 536 489.

Introduction

Fragment-based drug design (FBDD) is a mature and well established approach for drug discovery and optimization [4]. However, several limitations still exist related to the equipment and the expensive materials that are needed for the implementation of the protocols. For this reason chemoinformatics and computational tools can assist in parallel or in an independent manner to those discovery efforts by simplifying the fragment space to be explored or by pointing out which are the best spots within this space [5].

Although the definition of molecular fragment varies across the literature and depends on its intended use, the most common one takes into account size and physicochemical properties. This is the case of The Rule of Three definition [6], which states that the most successful fragments have a molecular weight of less than 300 kDa, a cLogP equal or less than 3 and a number of hydrogen bond donor and acceptor atoms of less than or equal to 3. This definition has been applied widely, but it should not be considered in absolute terms as some successful studies have employed fragments that do not fulfill one or even two of these recommendations [7].

Use of molecular fragments has some advantages from both experimental and computational standpoints. The number of fragments that can possibly exist is considerably lower than that of molecules with drug-like properties and fragments are smaller in size than most molecules in chemical libraries. In other words, fragment space is smaller that drug-like space and therefore it can be explored in a more detailed way while covering a larger diversity in the same amount of time. In some respects this can be viewed as if fragments compress the space described by drug-like molecules. However, as the size of a compound decreases its expected binding affinity towards a given target also decreases so that highly sensitive experimental techniques are then needed to detect such a weak binding event [5]. Even for computational techniques fragments challenge current algorithms and scoring functions, as these have been trained for use with drug-like molecules and are therefore unable to correctly determine interactions at the binding site or properly predict an accurate binding affinity value for a small fragment. For this reason, both the source of the fragments to be used and their physicochemical properties are of utmost importance for a protocol to be successfully used *in silico* on a given target [8].

Fragment databases are very diverse although they are commonly obtained from multiple suppliers which usually provide synthetically accessible small molecules or the result of filtering chemical libraries using the above mentioned rule of three. Another, and perhaps more interesting, alternative is to fragment the molecules present in drug-like databases into smaller pieces following a rational fragmentation scheme. One such popular and simple procedure was designed for drug analysis and classification of marketed compounds [9] whereas that known as RECAP (Retrosynthetic Combinatorial Analysis Procedure) was designed to address the issue of availability of high quality

building blocks for combinatorial chemistry [3]. The former method classifies molecular structures according to rings (cyclic fragments that form the base of the molecule), linkers (the acyclic parts that link the rings) and side chains (chemical groups attached to the rings) whereas the latter uses specific rules to disconnect certain parts of the molecules according to a list of simple chemical reactions that could yield the original compound. Accordingly, none of these methods was developed with a view to docking the resulting fragments into a protein target and evaluate their possible affinities computationally.

Fragment optimization methods need a yardstick to select the appropriate fragment each time and to warranty that the pathway chosen can be followed with no trouble. The most common parameter used it the ligand efficiency. Ligand efficiency indices (LEI) were introduced to normalize the binding free energy of compounds with different properties such as size, molecular weight, number of heavy atoms, etc. LEI [10] have demonstrated to be very effective in FBDD in both experimental and computational protocols [11] and have also been shown to properly describe the chemico-biological space (CBS) that is being used in fragment optimization [12]. Prospective and retrospective analyses [13; 14] have shown that a given path of optimization in CBS can be successfully predicted and followed using a LEIs framework and 2D planes.

In the following we describe a fully LEI-driven computational protocol that employs a succinct and diverse database of fragments and a growing scheme to suggest new target-oriented compounds with drug-like properties. The protocol encompasses a group of tools to perform binding site analysis, docking, scoring and LEI calculations. A Graphical User Interface (GUI) and some scripts were developed to ease its use and tested with two examples available from literature. In both cases, protein kinase B and thrombin, the protocol was able to identify the main features responsible for the binding of inhibitors and guided the process towards the more active molecules as found in the original studies.

Methods

Protein and ligand setup

Three-dimensional structures for proteins and ligands were prepared according to our standard reported protocol [15]. Briefly, proteins were extracted from Protein Data Bank entries upon removal of all other molecules in the file, assigned protonation states to all titratable residues and added hydrogen atoms and atomic charges using PDB2PQR [16] in accordance with the AMBER force-field [17]. Ligand protonation states were ascribed using OpenBabel [18] and molecules were stores as separate files. Fragments were built from SMILES strings using CORINA as the 3D molecular structure generator and stored in PDB file format as input for the protocol.

Fragmentation scheme

Our fragmentation scheme follows Murcko's [9] criteria and classifies any given molecular structure into three different entities: rings, linkers and side chains. However, we give to side chains the same relevance as to rings, considering both when generating fragments. Only linker-ring bonds are broken when fragments are being generated.

Two fragment databases were built, the first one starting from eMolecules [19] and using the modified Murcko's approach, which yielded 711,155 non-redundant fragments, and the second one, by defining a comprehensive number of common rings and attaching different substituents (chlorine, fluorine, methyl groups, acyclic linkers, amines, etc.) to the cyclic frameworks, in particular to those which are the most frequent ones found in the original work by Murcko et al. [9].

General overview

The protocol presented here is aimed at the design of drug-like molecules using a structure-based approach. It starts from a defined base fragment, uses an incremental construction algorithm and a scoring function to prioritize the most promising solutions, and is guided by the optimization of the LEI known as binding efficiency index (BEI) and surface efficiency index (SEI) for each candidate as a way to navigate the CBS efficiently [20].

To accomplish these tasks we developed a four-module program that allows the user to design new compounds starting from a desired scaffold. These modules perform the following jobs: (1) binding pocket analysis, to provide structural and energetic properties needed in subsequent steps; (2) base fragment placement according to the information retrieved from (1); (3) sampling, by adding new fragments to the base fragment using a growing algorithm; and (4) scoring, to evaluate the candidates using a scoring function that maximizes the square sum of BEI and SEI.

Binding pocket analysis

Our in-house binding pocket analysis tool (cGRILL), which is formally equivalent to Goodford's program GRID [21], evaluates selected cavities searching for affinity hotspots that are thought to contribute significantly to the binding free energy of putative small molecules.

The program reads in files in PQR or PDB format containing atomic coordinates and partial charges for each atom, which is characterized according to its connectivity (bond order), ring state, chemical type and non-bonded parameters using the Generalized AMBER Force Field (GAFF) [22]. Next, using a user-definable definition and resolution (spacing) of the search space, a cubic grid is built. At each grid point cGRILL evaluates the interaction energy between the whole receptor and five different probes combining van der Waals (Lennard-Jones potential), electrostatic (Coulombic) and hydrogen bonding [23] (geometry-based) interaction terms. These probes are

thought to summarize the main stereo-electronic properties of the binding pocket and are defined as follows: lipophilic (CH_3), hydrogen bond donor (H_4N^+) and hydrogen bond acceptor (=O), mixed hydrogen bond donor-acceptor (-OH), and hydrophobic. cGRILL implements the extended atom concept to simplify the probes and to improve the sampling speed over the target's surface [21].

The probes are reduced to their central atom with its partial charge increased depending on the atoms attached to it. Accordingly, the hydrogen bond acceptor probe has an assigned charge of -0.37e to better represents the partial negative potential on the oxygen atom when it is in a carbonyl group. On the contrary, the lipophilic C atom probe is neutral (charge 0e) and therefore only van der Waals interactions are calculated for it. The functions for hydrogen bond donor, acceptor and mixed donor-acceptor probes include an extra term (besides van der Waals and electrostatics) that accounts for the geometry of the hydrogen bond and depends on: (1) the distance between the acceptor and hydrogen atoms, (2) the angle between the donor, hydrogen and acceptor atoms, and (3) the relative orientation of the planes where the atomic orbitals of the acceptor and the hydrogen atoms are located. The hydrophobic probe is built on the lipophilic one but as an extra feature it adds the inverse of the default hydrogen bonding term. Thus, this probe will identify those regions where the interaction between the receptor and water molecules is unfavorable and the binding of a small molecule (or fragment) will improve by the displacement of any water molecules due to the hydrophobic effect [24].

After the mapping of the binding pocket is complete, the program filters out all those grid points with scores higher than a user-defined cutoff value (the interaction energy, by definition, is negative) for each probe, which is set by default to -12.0, -4.5, -7.0 and -1.7 kcal/mol for H_4N^+, =O, –OH, and hydrophobic probes, respectively. At each of the surviving points the probes compete according to their interaction energy values, and the best of the set becomes the representative probe at this grid point with its associated energy value. These grid points are then clustered and the local minima thus obtained are considered as hotspots and saved for further use. At each grid the clustering algorithm checks for the energy values of the nearest surviving points (within 2.0Å of distance) and if at least one of these points has a better value than its own the grid point is discarded [25].

Base fragment placement

The next step entails the selection of a starting or "base fragment" and its positioning within the binding pocket. This fragment typically comes either from an experimentally screened library coupled to crystallographic data to characterize its binding mode or from a previous virtual screening campaign.

An experimentally confirmed fragment hit at the starting location is not necessary but it increases the odds of a successful final design. Besides, and depending on the structure-activity landscape of the target, it is possible that the resulting optimized compounds will not share the binding mode of the starting fragment [26]. This risk can be minimized by using feature-rich fragments that establish relatively strong and defined interactions with the target. Computational starting points are also possible. However, docking-based poses are prone to very well-known errors [27] and do not always ensure that the final molecule will interact with the target as predicted. In the proof-of-concept applications presented here our methodology uses base fragments with experimentally determined binding modes.

Sampling and growing

This is the iterative process by means of which fragments from different databases are tried out (see below) and added to the base fragment. After each step a simplex method for energy minimization [28] can be used to refine the pose.

The growing algorithm employs four databases of putative fragments that are meant to match the four different types of hotspots detected by the binding pocket analysis tool, i.e. hydrogen bond donor, hydrogen bond acceptor, mixed hydrogen bond donor-acceptor and hydrophobic.

Sampling is a completely interactive step and starts with the user selecting an atom of the base fragment and a destination hotspot. Then, the program chooses the appropriate database of fragments depending on the type of hotspot and makes up candidate molecules by appending each fragment to the base fragment using hydrogen atoms from the former and the selected atom from the latter. Bond lengths are assigned according to the nature of the atoms being bonded. The SIMPLEX algorithm is then applied to fine-tune the pose at the binding pocket by optimizing: (1) the six rigid body degrees of freedom (translations and rotations) of the ligand, (2) the ligand rotatable bonds, and (3) the hydrogen bond donor groups of the target. Whether the base fragment remains frozen or it is allowed to move during this process is up to the user.

Scoring

We have implemented three different scoring functions that perform distinctly depending on the properties of ligand or fragment and the binding site: (1) MM-ISMSA [29], of general use; (2) ChemScore [30], better suited for non-charged ligands and hydrophobic pockets; and (3) HYDE [31], which works best for accurately placed ligands and could be applied to any protein–ligand (fragment) structure. The three scoring functions have been described extensively and tested against a wide range of targets [29; 32; 33] and scoring/ranking problems. This means that their relative strengths and weaknesses regarding binding pockets and ligands (fragments) are well known.

MM-ISMSA is an ultrafast and accurate force field-based scoring function that includes (1) a molecular mechanics (MM) part based on a 12−6 Lennard-Jones potential; (2) an electrostatic component based on an implicit solvent model (ISM) [34] with individual desolvation penalties for each partner in the protein−ligand (fragment) complex; and (3) a surface area (SA) contribution to account for the loss of water contacts upon protein−ligand (fragment) complex formation. As force field-based scoring functions are known to be well suited for pose prediction in docking and to discriminate efficiently amongst native and non-native candidates [35], MM-ISMSA is the default scoring function for sampling and final evaluation in our protocol.

The empirical function ChemScore decomposes the binding energy in terms of (1) a lipophilic contribution (only for non-polar atoms), (2) hydrogen bonding interactions (with a geometry-dependent function), (3) metal interactions (when present and only for hydrogen bond acceptor atoms), and (4) an entropic penalty for the freezing of rotatable bonds during the binding event (proportional to the number of rotatable bonds). In our implementation, this function lacks any penalty terms for high energy conformations or very tight binding molecules (atomic clashes). Therefore it will perform optimally in the evaluation of final poses that are force field-compliant in terms of both geometry and energy.

HYDE (HYdration and DEhydration) assumes that the main contributions to the binding free energy arise from hydrogen bonding interactions between the target and the ligand (fragment) and also that the accompanying desolvation event can either favor or penalize binding depending on the nature of the interacting chemical groups. The hydrophobic/hydrophilic nature of the atoms is determined by means of logP atomic contributions using empirically deduced coefficients from experimental values. The free energy is estimated depending on the balance between the geometry of the hydrogen bonds and the complementarity of target and ligand (fragment) surfaces.

To guide the growing scheme we have introduced a LEI-driven algorithm which calculates BEI and SEI for each candidate molecule and then plots the BEI vs. SEI efficiency plane [36]. To optimize both indices at the same time, the sum of their squared values is computed. This information will help the user to decide which of the best possible candidates will be selected as the base fragment for the next round of growing.

Results

The program

Our in-house pocket analysis tool, cGRILL, was implemented as a standalone C program making use of some parts of the molecular library presented in our previous works [37]. It can be used in command-line mode or within the molecular visualization

and editing program PyMOL [38] as a graphical user interface (GUI) plugin. The GUI has four different tabs (Fig. 1): 1) *Run cGRILL*, the interface to define the search space at the protein and the grid spacing, the name of the protein file and two buttons to start and to stop the calculation; 2) *Load Grids*, to load the grids into the PyMOL session from either the current or a previously saved analysis, and to display the calculated affinity maps with an arbitrary at user-definable cutoff values; 3) *Configuration tab*, to inform the program where the binary of the cGRILL code is located and the working directory to store the results; and 4) *About tab*, where credits and support information are provided.

Figure 1 near here

The main protocol interface was developed as a python script that processes the base fragment and the specified fragment library and builds up new molecules by using a linker program (developed in C). This program generates all the possible combinations between the two entities, the base fragment and the library. Then, the script calls an energy minimization and sampling routine that processes the protein and each putative molecule.

The last part of the protocol is driven by another script which, using the OpenBabel Python wrapper Pybel [39] computes the polar SA and the molecular weight of the candidates to calculate BEI and SEI values and the efficiency plane, being ready for the next step, where the user can chose its starting point.

Example 1: Protein Kinase B (PKB) inhibitors with a 4-phenyl-1*H*-pyrazol scaffold

PKB is a serine/threonine kinase that regulates many pathways in cell growth and differentiation. Saxty *et al.* published a series of compounds obtained from a fragment-based lead discovery campaign, [40] and provided inhibitory activity values (IC_{50}), a simple ligand efficiency metric and the crystal structure of each intermediate in the optimization process. Using all this data we applied our protocol with the aim of comparing *in silico* proposed candidates (structures and binding modes) with the experimental results just mentioned and their specific location on the efficiency plane.

After their fragment screening the authors identified 5-methyl-4-phenyl-1*H*-pyrazole (Fig. 2) as a hit scaffold. This small molecule was soaked into a crystal of PKA-PKB hybrid proteins and the binding mode was determined. Starting from this X-ray structure (PDB id. 2UW3) we selected this moiety as the base fragment and the protein without the ligand was submitted to binding pocket analysis.

Figure 2 near here

Three different hotspot regions were identified (Fig. 3): (1) a strong hydrophobic cluster, located at the same place as the phenyl ring of the base fragment and two mixed hydrogen bond donor-acceptor hotspots which are coincident with the two nitrogen atoms in the pyrazole ring; (2) right at the top of (1) a mixed hydrogen bond donor-

acceptor spot spreading along the bottom of the pocket pointing towards a cavity (hereafter referred as the next step in the optimization pathway); and (3) a positively charged cluster with a mixed hydrogen bond donor-acceptor character located near the DFG motif, a key element in kinases activation/inactivation processes [41].

Figure 3 near here

Next, we started the growing process using cluster (3), the *para* position of the phenyl ring, because cluster (1) was already in use by the base fragment and cluster (2) is farther away from the initial fragment. Each possible combination of fusing individual fragments from the library and the base fragment was explored. Once a putative molecule was built, the sampler module optimized the ligand's rotatable bonds while keeping the base fragment rigid and used the GAFF non-bonded terms to select the best pose. Finally, at the end of the first round, SEI and BEI were evaluated for all those compounds giving rise to a favorable binding free energy using their scores and plotted for visual inspection (Fig. 4). Table 1 shows the top four compounds selected together with their BEI and SEI associated values.

Figure 4 near here

Table 1 near here

Example 2: Highly potent thrombin inhibitors

Thrombin is a serine protease involved in the blood coagulation cascade and therefore a main target for anti-clotting agents. Despite the fact that many compounds have been developed only a few have made it into the clinic. Here we follow the optimization of a series of inhibitors originally published by Klebe *et al.* [42] and based on the structure of the peptide inhibitor D-Phe-Pro-Arg (Fig. 5, left). The authors explored S1 and S3 subpockets and rationalized the activity enhancements as they went along the series (Fig. 5, right).

Fig. 5 near here

The binding pocket analysis step of our protocol started with the X-ray structure of thrombin (PDB id. 2ZFP) co-crystalized with 1-[(2R)-2-aminobutanoyl]-N-(3-chlorobenzyl)-L-prolinamide, a close analog of the peptide inhibitor whose scaffold was used as the base fragment for further optimization attempts. cGRILL identified three main putative interaction areas (Fig. 6): (1) the S1 subpocket, where a cluster of three different hotspots was found. First, a hydrophobic one in the middle of the cavity (a ring would clearly contribute to the binding free energy); a positively charged hotspot at one end and a mixed hydrogen bond donor-acceptor region at the other end of the cavity. Therefore, a positively charged group in the former (*para* position) and a halogen atom in the latter (*meta* position, *m*-) would provide candidates with increased affinity; (2) the central part of the binding pocket, filled with four hotspots of mixed hydrogen bond donor-acceptor character and a positively charged region; and (3) the S3 subpocket,

with a single hydrophobic hotspot. Selecting the scaffold derived from the tripeptide (Fig. 5) as the base fragment (extracted from the previously described co-crystallized structure) there are two different optimization paths that can be taken: the S1 and the S3 subpockets. Given that in both cases the hotspot nearest to the base fragment is hydrophobic we chose to use a ring fragment library composed of saturated and unsaturated rings with 3 to 6 atoms including common combinations of nitrogen, oxygen and sulphur atoms with usual substituents (e.g. halogens, methyl, amines). Each possible link between the base fragment and the additional fragments were explored at *meta*, *para*, and *ortho* positions.

Figure 6 near here

The results on S1 confirm that the *m*-chlorine aromatic rings of 5 and 6 members are the preferred ones and that the chlorine atom is well buried into the pocket. In addition, positively charged groups provide the largest increase in the binding free energy due to an extra interaction with residue Asp189 at the bottom of the S1 pocket. On the other hand, for S3 only pure hydrophobic fragments were used as there was no polar hotspot near the original hydrophobic core. The results show that the most efficient compounds are those with saturated rings of 5 to 6 carbon atoms and mono-substituted ciclopentane rings with a bromide atom filling the lower part of the subpocket. SEI and BEI were then evaluated for the top compounds using their scores and plotted for visual inspection (Fig. 7). Table 2 shows the top compounds selected for S1 with their BEI and SEI values.

Figure 7 near here

Table 2 near here

Discussion

Our novel FBDD protocol uses (1) a database of diverse and limited number of fragments obtained from common synthesizable scaffolds and (2) an incremental growing scheme where molecular growth is driven by efficiency planes employing BEI and SEI. In this way we ensure that the new proposed candidates will be synthetically accessible and endowed with the appropriate physicochemical properties. These two issues are, in essence, the main differences among the many programs that have been developed since LUDI, the pioneer in the field, was published back in the early 1990s [1].

On the one hand, the exploration of the fragments' chemical space should be done in a way that properly covers the largest possible percentage of it while, at the same time, keeps the computational costs affordable and produces molecules that could be actually synthesized by chemists. Some strategies that follow on these guidelines are Fragment Optimized Growth [43], whose growing engine uses a Markov chain to bias the search

process; COLIBREE [44], which makes use of a particle swarm optimization algorithm and a series of linkers related to certain chemical reactions to ensure synthetic accessibility; RECAP [3], which follows chemically inspired rules to break molecular databases into fragments that may be connected to the main scaffold afterwards through a feasible reaction; and SQUIRRELnovo [45] a molecular superposition algorithm using bioisosters and Flux [46] a program that employs a stochastic algorithm and a ligand-based similarity score to a given template.

On the other hand, highly scored candidates must fulfill certain physicochemical properties to be properly considered as hits. Most implementations rank molecules by relying on a fast scoring function coming from docking programs instead of using a more reliable function, although more expensive in computational terms, such as MM-PBSA [47], Free Energy Perturbation (FEP) [48], or Thermodynamic Integration (TI) [49]. The main problem with this approach arises from the fact that key properties such as solubility or size are not adequately considered. Therefore, these functions tend to award high scores to large molecules (as a consequence of its pairwise-interaction nature) irrespective of the goodness of their fits. Two examples of mixing different approaches to guide the process of selecting compounds are LigBuilder 2.0 [50], where a genetic algorithm couples a binding affinity prediction and the evaluation of physicochemical properties, and PhDD [51], which uses non-compliance to Lipinski's rule of 5 and an excessive number of rotatable bonds as criteria to discard compounds with a poor profile. However, Lipinski's rules or another drug-like set of properties may be too restrictive a criterion and lead to discarding some promising molecules.

Taking into account the successful results obtained for the two tested targets and bearing in mind that BEI and SEI rely on estimated force-field based interaction energies, the protocol developed here by coupling a growing algorithm and a LEI-driven scheme seems to be an appropriate way to optimally navigate through the chemicobiological space finally yielding the most active compounds in both cases. However, care must be taken as the protocol not only depends on the selection and positioning of the base fragment but also on how accurate we describe de binding site. Summarizing, computational schemes as the one presented here pave the way to a more rational drug design paradigm based on the incremental construction of putative candidates.

Conclusions

A group of computational tools to perform in silico FBDD has been developed using an innovative approach to solve the two most important problems found in this techniques, namely (1) the synthetic accessibility of the new molecules by making them up from databases of high frequently fragments used in marketed compounds, and (2) the guidance of the optimization process by the simultaneous fine-tuning of two ligand efficiency indices (BEI and SEI). By combining these two approaches we were able to

find the most promising compounds in two retrospective examples using protein kinase B and thrombin as targets.

CGILL and the scripts are open source and can be downloaded free of charge following registration at the CBM Bioinformatics Unit's web page (http://ub.cbm.uam.es/).

Acknowledgement

This work was supported by grants from CICYT (SAF2009-13914-C02-02 to F.G.) and Comunidad Autónoma de Madrid (S-BIO-0214-2006 [BIPEDD] and S2010-BMD-2457 [BIPEDD2] to A.M. and F.G.). A.M. acknowledge financial support from Fundación Severo Ochoa through the AMAROUTO program. A.C.C. is the recipient of FPU grant AP2009-0203 from the Ministerio de Educación.

References

[1] H.-J. Böhm, Journal of computer-aided molecular design, 6 (1992) 61.
[2] C.A. Lipinski, F. Lombardo, B.W. Dominy and P.J. Feeney, Advanced drug delivery reviews, 23 (1997) 3.
[3] X.Q. Lewell, D.B. Judd, S.P. Watson and M.M. Hann, Journal of chemical information and computer sciences, 38 (1998) 511.
[4] C.W. Murray and D.C. Rees, Nature chemistry, 1 (2009) 187.
[5] M. Congreve, G. Chessari, D. Tisi and A.J. Woodhead, Journal of medicinal chemistry, 51 (2008) 3661.
[6] M. Congreve, R. Carr, C. Murray and H. Jhoti, Drug discovery today, 8 (2003) 876.
[7] H. Jhoti, G. Williams, D.C. Rees and C.W. Murray, Nature reviews Drug discovery, 12 (2013) 644.
[8] C. Sheng and W. Zhang, Medicinal Research Reviews (2012).
[9] G.W. Bemis and M.A. Murcko, Journal of medicinal chemistry, 39 (1996) 2887.
[10] C.H. Reynolds, B.A. Tounge and S.D. Bembenek, Journal of medicinal chemistry, 51 (2008) 2432.
[11] S.D. Bembenek, B.A. Tounge and C.H. Reynolds, Drug discovery today, 14 (2009) 278.
[12] C. Abad-Zapatero and D. Blasi, Molecular Informatics, 30 (2011) 122.
[13] D. Blasi, G. Arsequell, G. Valencia, J. Nieto, A. Planas, M. Pinto, N.B. Centeno, C. Abad-Zapatero and J. Quintana, Molecular Informatics, 30 (2011) 161.
[14] D. Tanaka, Y. Tsuda, T. Shiyama, T. Nishimura, N. Chiyo, Y. Tominaga, N. Sawada, T. Mimoto and N. Kusunose, Journal of medicinal chemistry, 54 (2010) 851.
[15] Á.C. Cabrera, R. Gil-Redondo, A. Perona, F. Gago and A. Morreale, Journal of computer-aided molecular design, 25 (2011) 813.
[16] T.J. Dolinsky, J.E. Nielsen, J.A. McCammon and N.A. Baker, Nucleic acids research, 32 (2004) W665.
[17] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell and P.A. Kollman, Journal of the American Chemical Society, 117 (1995) 5179.

[18] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch and G.R. Hutchison, Journal of cheminformatics, 3 1.
[19] J.J. Irwin and B.K. Shoichet, Journal of chemical information and modeling, 45 (2005) 177.
[20] C. Abad-Zapatero, O. Perisic, J. Wass, A.P. Bento, J. Overington, B. Al-Lazikani and M.E. Johnson, Drug discovery today, 15 (2010) 804.
[21] P.J. Goodford, Journal of medicinal chemistry, 28 (1985) 849.
[22] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman and D.A. Case, Journal of computational chemistry, 25 (2004) 1157.
[23] D.N. Boobbyer, P.J. Goodford, P.M. McWhinnie and R.C. Wade, Journal of medicinal chemistry, 32 (1989) 1083.
[24] C. Tanford, Science, 200 (1978) 1012.
[25] J. Ruppert, W. Welch and A.N. Jain, Protein Science, 6 (1997) 524.
[26] R. Brenk, L. Naerum, U. GrÃ¤dler, H.-D. Gerber, G.A. Garcia, K. Reuter, M.T. Stubbs and G. Klebe, Journal of medicinal chemistry, 46 (2003) 1133.
[27] J.B. Cross, D.C. Thompson, B.K. Rai, J.C. Baber, K.Y. Fan, Y. Hu and C. Humblet, Journal of chemical information and modeling, 49 (2009) 1455.
[28] J.A. Nelder and R. Mead, The computer journal, 7 (1965) 308.
[29] J. Klett, A. Nuñez-Salgado, H.G. Dos Santos, Á. Cortés-Cabrera, A. Perona, R. Gil-Redondo, D. Abia, F. Gago and A. Morreale, Journal of Chemical Theory and Computation, 8 (2012) 3395.
[30] M.D. Eldridge, C.W. Murray, T.R. Auton, G.V. Paolini and R.P. Mee, Journal of computer-aided molecular design, 11 (1997) 425.
[31] N. Schneider, G. Lange, S. Hindle, R. Klein and M. Rarey, Journal of computer-aided molecular design, 27 (2013) 15.
[32] G.L. Warren, C.W. Andrews, A.-M. Capelli, B. Clarke, J. LaLonde, M.H. Lambert, M. Lindvall, N. Nevins, S.F. Semus and S. Senger, Journal of medicinal chemistry, 49 (2006) 5912.
[33] I. Reulecke, G. Lange, J.r. Albrecht, R. Klein and M. Rarey, ChemMedChem, 3 (2008) 885.
[34] A. Morreale, R. Gil-Redondo and A.R. Ortiz, PROTEINS: Structure, Function, and Bioinformatics, 67 (2007) 606.
[35] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley and J.K. Perry, Journal of medicinal chemistry, 47 (2004) 1739.
[36] C. Abad-Zapatero, (2007).
[37] Á.C. Cabrera, J. Klett, H. G. Dos Santos, A. Perona, R. Gil-Redondo, S.M. Francis, E.M. Priego, F. Gago and A. Morreale, Journal of chemical information and modeling, 52 (2012) 2300.
[38] L. Schrodinger, See http://pymol.org.
[39] N.M. O'Boyle, C. Morley and G.R. Hutchison, Chem Cent J, 2 (2008).
[40] G. Saxty, S.J. Woodhead, V. Berdini, T.G. Davies, M.L. Verdonk, P.G. Wyatt, R.G. Boyle, D. Barford, R. Downham and M.D. Garrett, Journal of medicinal chemistry, 50 (2007) 2293.
[41] D.K. Treiber and N.P. Shah, Chemistry & Biology, 20 (2013) 745.
[42] B. Baum, L. Muley, A. Heine, M. Smolinski, D. Hangauer and G. Klebe, Journal of molecular biology, 391 (2009) 552.
[43] P.S. Kutchukian, D. Lou and E.I. Shakhnovich, Journal of chemical information and modeling, 49 (2009) 1630.
[44] M. Boehm, T.-Y. Wu, H. Claussen and C. Lemmen, Journal of medicinal chemistry, 51 (2008) 2468.
[45] E. Proschak, H. Zettl, Y. Tanrikulu, M. Weisel, J.M. Kriegl, O. Rau, M. Schubertâ€Zsilavecz and G. Schneider, ChemMedChem, 4 (2009) 41.

[46] U. Fechner and G. Schneider, Journal of chemical information and modeling, 46 (2006) 699.
[47] I. Massova and P.A. Kollman, Perspectives in drug discovery and design, 18 (2000) 113.
[48] S.N. Rao, U.C. Singh, P.A. Bash and P.A. Kollman, Nature, 328 (1987) 551.
[49] K.-W. Wu, P.-C. Chen, J. Wang and Y.-C. Sun, Journal of computer-aided molecular design, 26 (2012) 1159.
[50] Y. Yuan, J. Pei and L. Lai, Journal of chemical information and modeling, 51 (2011) 1083.
[51] Q. Huang, L.-L. Li and S.-Y. Yang, Journal of Molecular Graphics and Modelling, 28 (2010) 775.

Table 1. Four top molecules resulting from the first optimization round in the search of PKB inhibitors: chemical structure, BEI and SEI LEIs values and some comments.

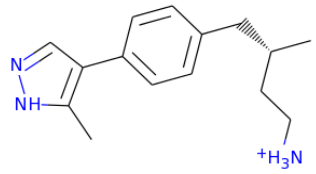
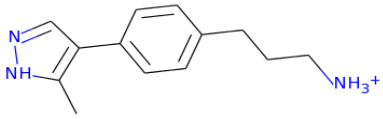
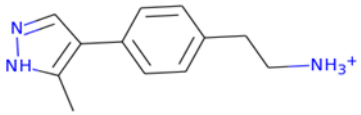
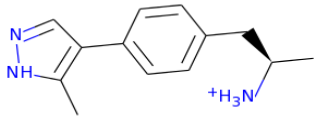
Compound	BEI	SEI	Comment
	55.2	7.2	This compound is pointing towards the next step of the optimization pathway.
	53.1	7.8	Most efficient compound found experimentally (IC_{50} = 3.0 ± 1.2 μM, LE=0.51)
	55.1	7.5	Compound Selected by the authors, although not the most efficient (IC_{50} = 5.2 ± 3.3 μM, LE=0.48)
	50.5	8.0	This compound is pointing out the next step of the optimization pathway.

Table 2. Top molecules resulting from the first optimization round in the search of thrombine inhibitors: chemical structure and BEI and SEI LEIs values.

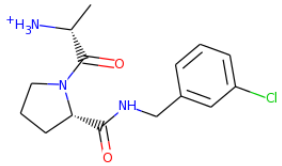
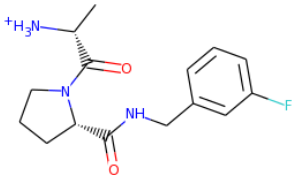
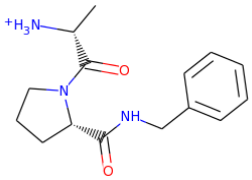
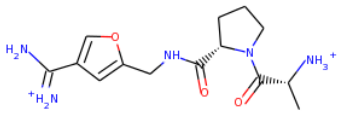
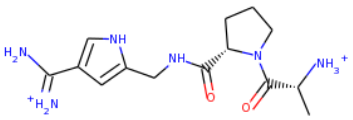
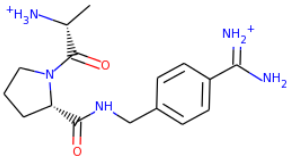
Compound	BEI	SEI
S1[1]		
	19.4	10.7
	18.4	10.2
	19.1	10.0
S1[2]		
	52.1	13.7
	47.9	12.2
	38.0	11.6

Figure 1. cGRILL graphical user interface.

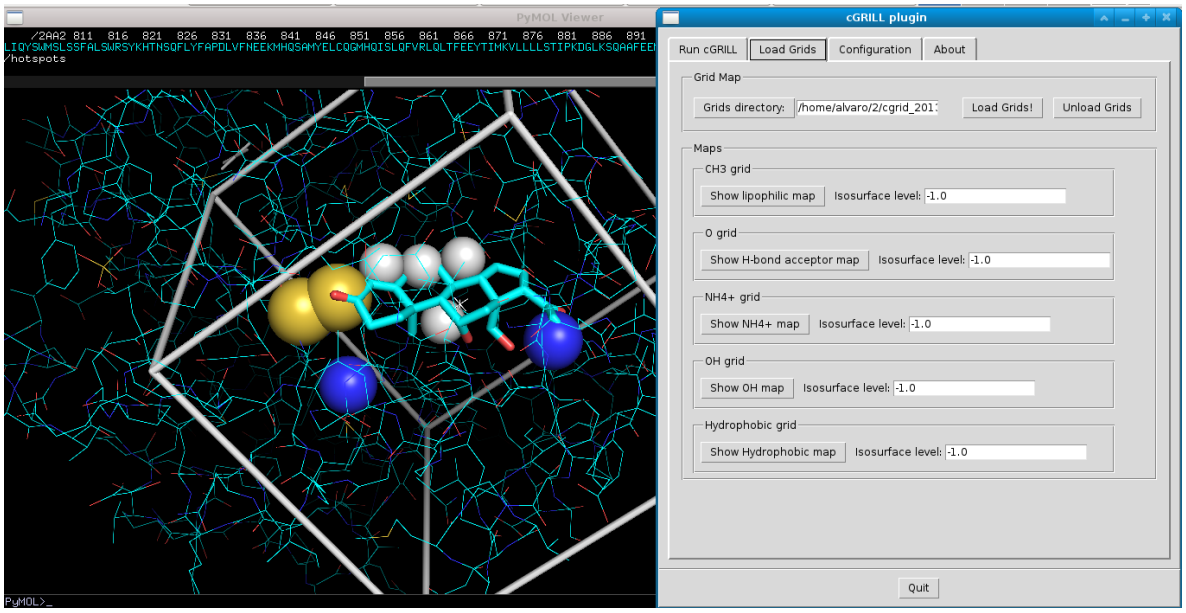
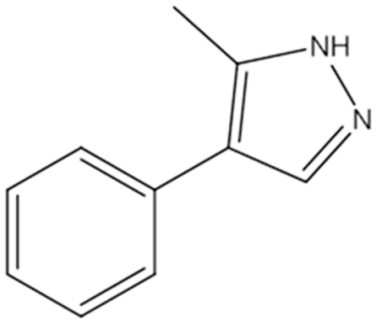


Figure 2. Base fragment for the optimization process.



5-methyl-4-phenyl-1H-pyrazol

Figure 3. Binding site analysis results for the kinase PKA.

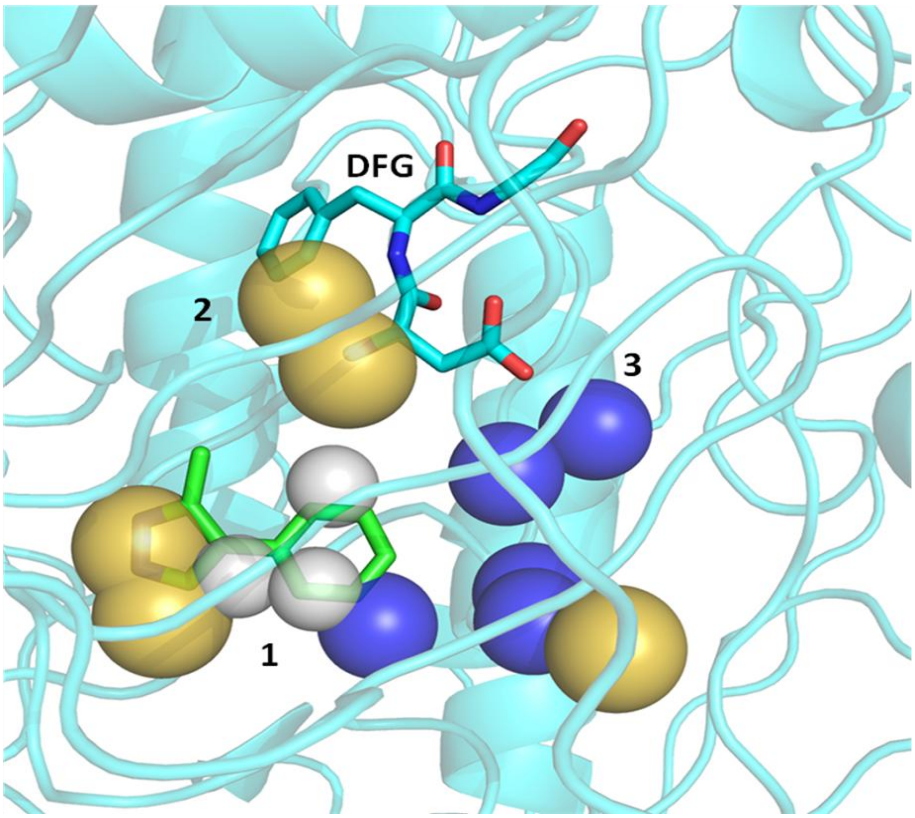


Figure 4. Efficiency plane for the putative kinase inhibitors

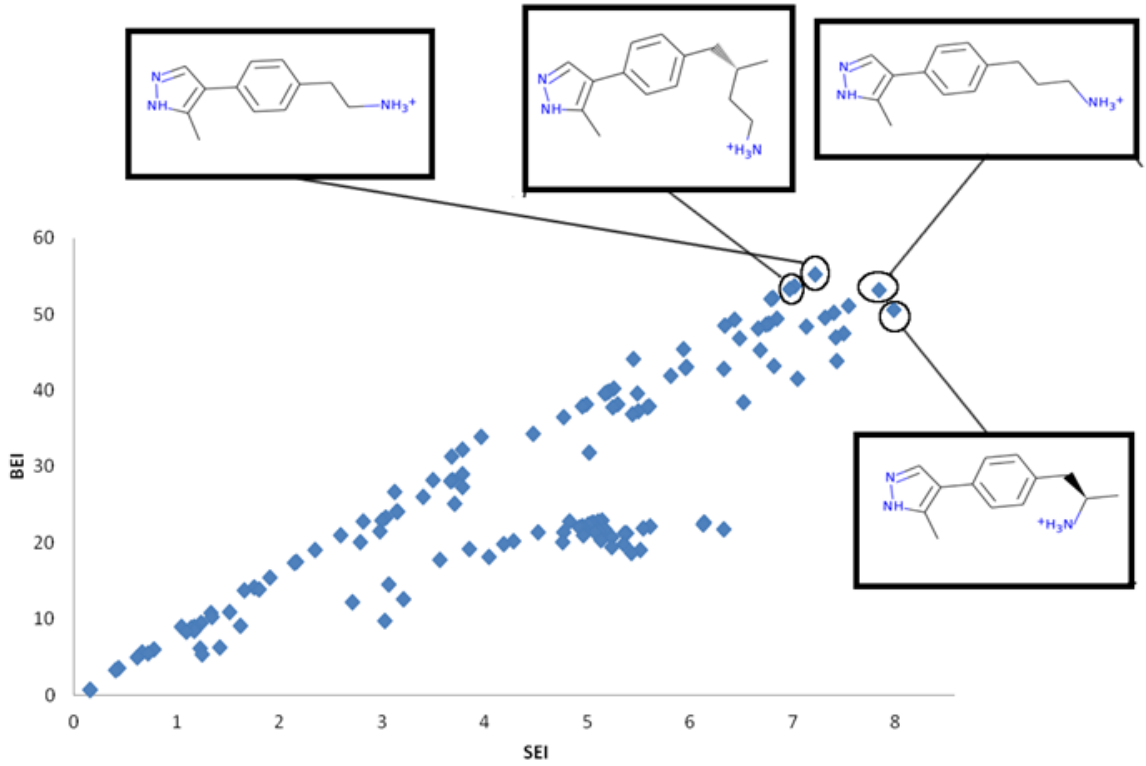


Figure 5. Base fragment for thrombin and the optimization pockets.

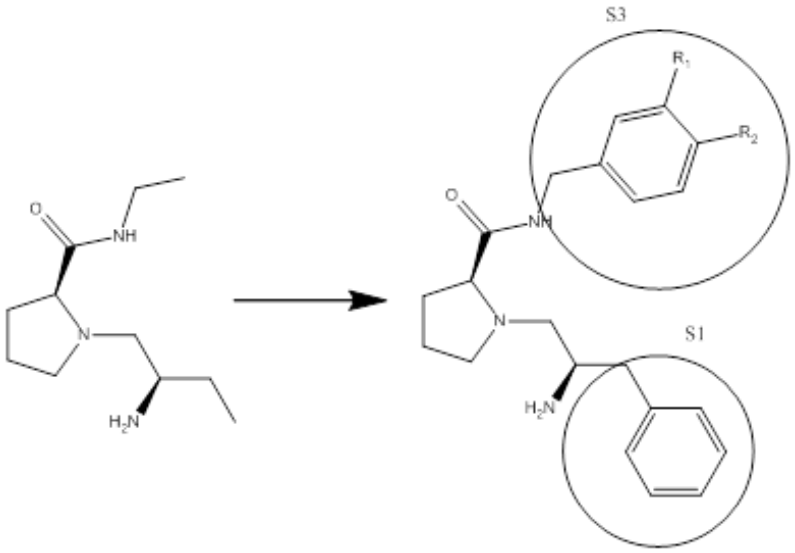


Figure 6. Thrombin binding site analysis results.

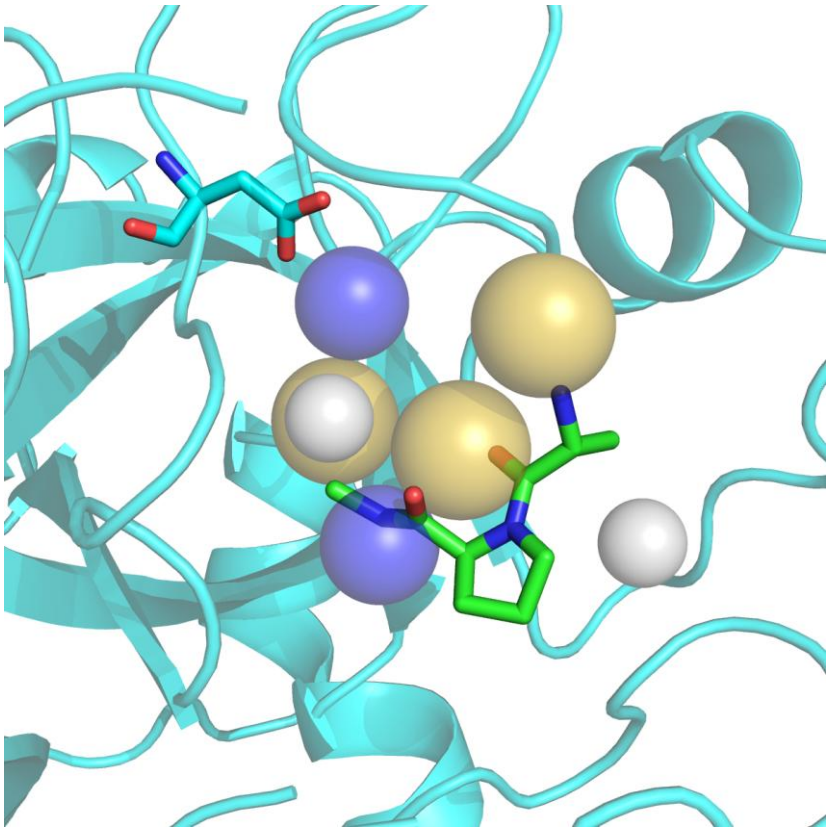
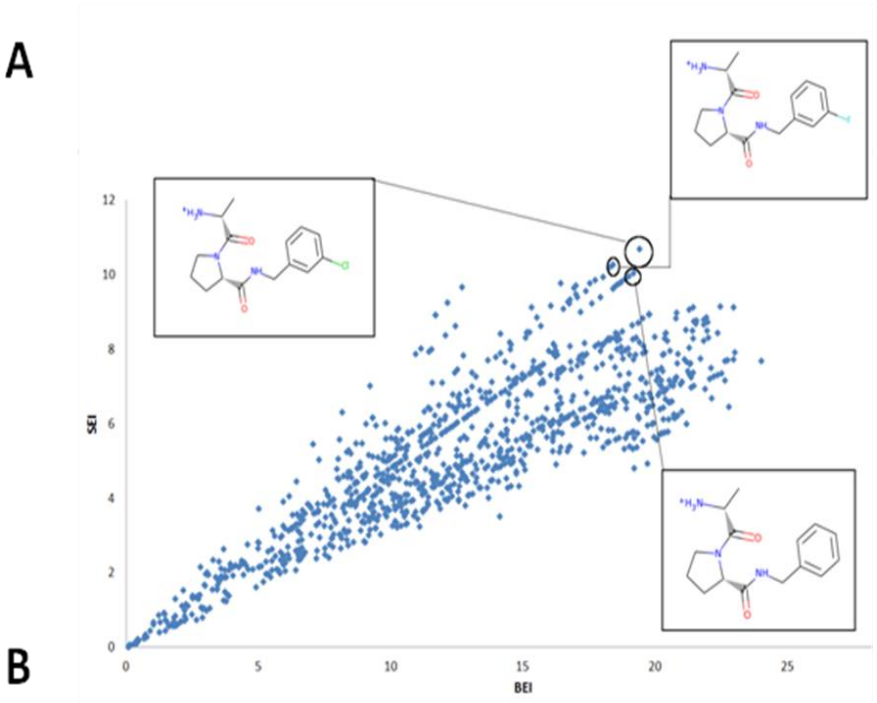
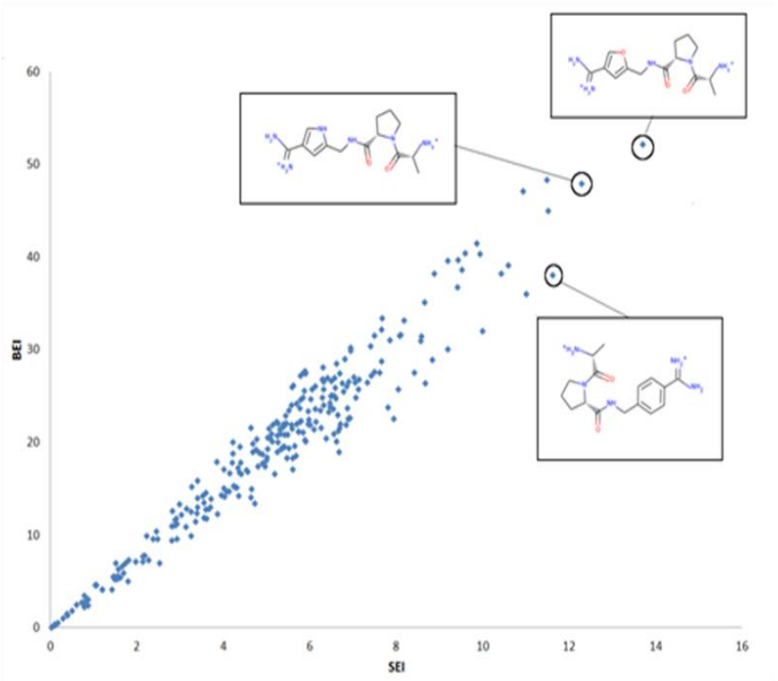


Figure 7. Efficiency planes for S1 subpocket analysis with normal (A) and positively charged fragments (B).





5. DISCUSIÓN

The computational study of pharmacological space is a daunting task because it entails not only the intermolecular interactions governing the binding process, which must be modeled using atomistic scales and rigorous physical theories, but also the thousands of molecular relationships in the biological systems where drugs work which should be modeled using large networks containing all the relevant information.

In this thesis, a few research items have already been presented that, at least in part, try to address these topics to some extent: i) articles I, II and VI are related to low-level drug discovery approaches and to the introduction of these capabilities in an integrated platform for drug discovery and ii) articles III, IV and V introduce a more general concept where the high-dimensional spaces are studied and their complexity is analyzed or reduced for prediction of drug activities, side effects or compounds optimization.

We started off by addressing the main issues of the drug discovery platform that existed in our lab at that time. The improvement in “ease of use” was accomplished by implementing a new graphical user interface as a plug-in for the popular molecular visualization and editor program PyMOL (Schrodinger). The user introduces a few parameters regarding the system under study and the interface re-uses this data to generate complex configuration files and set up execution environments for the underlying applications. In this new version 1.5 of the VSDMIP (Cabrera, Gil-Redondo et al. 2011), a base system for ligand-based virtual screening was also developed where a 2D or 3D ligand similarity query could be executed using a cartridge developed to run inside the database. Chemical group filtering and obtention of fingerprints were made possible by using SMARTS patterns in the groups and OpenBabel MACCS files (O'Boyle, Banck et al.). The new SMARTS definitions for the 2D pharmacophoric fingerprints called CATS (Schneider, Neidhart et al. 1999), were implemented and tested. The 3D pharmacophores, initially integrated in the database, were redesigned for scalability as a new distributable module.

Despite these improvements, the platform still suffers from some limitations. The MySQL database, which is also a distinct feature that allows storing the results and pre-processed molecules for different protocols, currently imposes some restraints on the flexibility of the molecules. Degrees of freedom of small molecules are reduced to rigid-body motions and discrete torsional freedom. All torsionals are encoded as a string of possible dihedral angles corresponding to minima and the values selected for a current conformer. This approach clearly limits the true flexibility of the ligands, where the movements of the rotatable bonds could strongly influence the positions of the atoms, producing *too rigid* models that do not correspond to the poses derived experimentally. On the other hand, this way of saving the information reduces significantly the storage costs and accelerates VS protocols because ensembles of pre-generated conformers are utilized.

In the last 10 years, the generic workflow platforms KNIME (Berthold, Cebron et al. 2008) and Pipeline Pilot (Pilot) have gained ground and limited the development of custom-made alternatives. These platforms present some advantages that make them preferable to develop in-house technology: large communities of users and developers exist, strong companies support them, third parties develop plug-ins and configuration is easily accomplished through graphical interfaces. Possible future developments of VSDMIP should address these facts and try to use standard platforms like the open source KNIME as a modern base to deploy the VS protocols.

The speed and accuracy problems in CDOCK (Pérez and Ortiz 2001) were undertaken by designing and implementing a new flexible docking tool called CRDOCK (Cabrera, Klett et al. 2012). This new program was implemented from scratch in C language as a series of library routines that allowed its faster development and modularization. We have licensed the docking program under the GNU General Public License (GPL) v2 or above to make the code available to the scientific community. CRDOCK improved the runtime speed by simplifying the search space, pre-scanning the binding pocket with three probes (hydrogen bond acceptor, hydrogen bond donor and lipophilic). In this manner, the compounds' features are placed only where chemical complementarily could be expected. When no distinct features can be found in the molecules or in the binding pocket, the program relies on a traditional Monte Carlo Simulated Annealing algorithm or performs an exhaustive search. The tool is built on top of a simple library of molecular operations which is also the base of other tools in this thesis. The CDOCK behavior could be emulated thanks to this modularity. The problem with the scoring function was approached by acknowledging that no universal scoring function exists that performs well in all the biological systems of interest. For this reason, several scoring function were developed and tested, selecting the general CRScore function, based on GlideScore (Friesner, Banks et al. 2004), by default but allowing the user to select others with a simple flag.

An improvement in pose prediction success is obtained as a result of the methodological changes. For rigid docking the program achieves a 91% (similar to the original CDOCK), but it obtains a reasonable 74% when flexibly modeled ligands are used instead. Far from perfect, this new version inherits some of the design faults of CDOCK and generates some new problems of its own. First, it uses the same approach to ligand flexibility than did CDOCK: a pre-generated ensemble of conformers. This design responds to the need of high speed in the calculations but reduces the internal sampling of the ligand. Besides, algorithms and near-linear runtimes routines for distributing the code over several processors are easier to implement. The main cost is that, as stated in the manuscript, this procedure is strongly dependent on the conformer generator and the quality of the models produced by the 1D-to-3D converter [this weakness is shared with all docking programs

(Friesner, Banks et al. 2004)], and, in many cases, near-native poses could not be located because of a defective ensemble of conformations. This explains some of the failures when results from flexible and rigid docking are compared. Another source of failure originates from the fact that other binding modes compatible with the electron distribution or electrostatic potential of the ligand are also possible when different conformations are used and the scoring function may *fail* to identify the one modeled by nature (Neves, Totrov et al. 2012).

Scoring function problems are solved partially by the new two-step scoring mechanism. The criteria in the sampling and final steps to select a *proper* solution are different. In the first step, we sample for possible poses, *i.e.* with a reasonable geometry, lack of clashes with the target and no feature mismatches. All of these are achieved through the use of the non-bonding terms of the Generalized AMBER force-field potential (Wang, Wolf et al. 2004). Force field-based scoring functions have been recognized as very competent at selecting *good* poses but not very talented at selecting near-native poses due to its strict definition of the 12-6 van der Waals term (Friesner, Banks et al. 2004). For that reason, in the final step, all selected poses based on this first score are re-scored with a simplified hybrid potential, CRScore, that includes the scaled electrostatic and van der Waals terms plus hydrogen bonding and hydrophobic terms from ChemScore (Eldridge, Murray et al. 1997).

Desolvation is still an unsolved issue in docking programs despite the fact that most of them include some sort of algorithm to account for this molecular recognition event. The CDOCK scoring function adds desolvation energies by using the Implicit Solvent Model (Morreale, Gil-Redondo et al. 2007) developed by Morreale *et al*. ISM solved some problems that are common to other docking programs by accounting for the tendency to bury ligand charged groups inside the target. However, the penalties for mismatched polar groups were greatly underestimated. Attempts were made to add this ISM term to the CRScore function and use the PDBBind (Wang, Fang et al. 2004) set of X-ray structures with affinities measured for calibration and testing but the resulting potential tended to be overzealous regarding the desolvation costs. More work is underway to adapt other semi-explicit scoring functions such as HYDE (Schneider, Lange et al. 2013) or ChemPLP (Korb, Stutzle et al. 2009) in order to improve the current CRScore-ISM scoring function.

In the last work on classical CADD methods (article III) we introduced a method to compress the search space in docking experiments. First of all, we validated the performance of the CDOCK (Pérez and Ortiz 2001) program when using small molecules. Fragments, due to their size and nature, may not hold distinctive moieties such as hydrogen bond donors/acceptors or ionizable groups and this may cause problems when

the docking program and its scoring function attempt to find the most suitable pose (Sándor, Kiss et al. 2010). This is the main reason why they are described as more challenging for the current tools than the larger, more drug-like molecules that developers had originally in mind. In this case, the *hard* scoring function implemented in CDOCK appears to be an advantage to dock these small molecules since the minima are sharper. Bearing in mind that we seek to reduce the amount of time spent in VS protocols, we have compressed the chemical space by selecting a subset of small molecules or fragments that are present in ZINC (Irwin and Shoichet 2005)]. and available for purchase. Since the synthetic accessibility of the compounds in a drug-like chemical space may depend on the small pieces available to the chemists, we estimated that an ensemble of *circa* 2500 molecules was reasonable to map this entire space. The main advantage of this method is that it avoids the clustering, and more importantly, the construction of an all-versus-all matrix needed for that process (Varin, Bureau et al. 2009) by using a stochastic algorithm that ensures the non-redundancy of the selected fragments. Comparing the results from a standard protocol to those obtained by using the compressed-expanded libraries presented here, we concluded that the number of ligands to dock was successfully reduced while preserving enough chemical diversity to achieve equivalent performance on a large variety of targets. However, we found some limitations in the process due to the nature of the ligands in our test sets. This is the case of Nuclear Hormone Receptors (NHR), whose typical ligands contain a steroid scaffold that was not available in our fragment set. This fact only highlights the need to prepare carefully a specific target-driven chemical library to achieve the best results.

Setting aside the atomistic approach of the classical CADD employed in this thesis, two new global methods were developed to address the drug discovery problem from a different perspective. In the first work (article IV), we validated the use of ElectroShape 3D descriptors to construct polypharmocology networks and to predict side effects on the basis of information present in a pharmacology database such as DrugBank. We measured the amount of useful information of these networks and compared to the ones built from aligning targets binding sites. In addition, we used all this information to predict possible side effects based on the profiles of interacting targets. Finally we did a retrospective on-target and off-target validation of the networks using several well known datasets such as the Directory of Useful decoys or the WOMBAT sets. Also, we have used a small set of non-included activities for the off-target prediction with considerable success. Regarding the side effects, mild results were obtained. Apparently, the tool is able to discard false side effects from the database; however, it has major problems to recognize most of the reported effects. It should be pointed out that many of these effects could not be related to only one target or the description is too vague to be correlated with a single source. As a

last contribution, we built a very simple web application that allows any user to perform unlimited queries against the DrugBank-based polypharmacology network. This tool predicts target profiles and side effects using ElectroShape 3D descriptors or the 2D Morgan fingerprints. It also represents the relationships between a selected target and their neighbors in the ligand and target spaces and shows the related compounds in the database with direct links to DrugBank.

It is clear that there is room for improvement. The main limitation of this approach is that the predictions are entirely based on the knowledge available in the database. In this case we used DrugBank, which is a manually curated database. However, we found several problems due to the inclusion of very general and non-specific ligand-target pairs and the poor discrimination between cofactors, natural ligands and drugs for human use. Attempts to employ another database such as ChEMBL solved some of these issues but incorporated new problems of their own, derived from the cut-off of binding energy introduced to consider any compound as a true binder. Side effects are poorly predicted in this version, probably due to a limited compilation of target-side effects relationships and to the fact that more complex mechanisms than those reported here are responsible for these problems. The web application is very limited, and it only allows the representation of the first layer of neighbors of any target. This possibly undermines one of the most important uses of this network-based platform which is the investigation of hidden relationships. The results are also difficult to store and the interface is sometimes oversimplified. These problems may be addressed in future developments.

In the penultimate piece of work, the AtlasCBS concept was materialized in the shape of a web tool that is able to describe the chemico-biological space in an effective manner. It uses the LEI framework and 2D planes to achieve such a description. Users can not only explore the contents of the most popular CBS databases but also upload their personal projects in a simple semicolon separated file or introduce the information manually via the web page. The display of hard-coded data from public sources together with the personal information provided becomes a powerful tool to analyze and compare projects and spaces. Registration is required for uploading a private dataset in order to ensure certain privacy to the users. The server implementation is based on open source technology such as the Chemistry Development kit (CDK), the Apache Tomcat application server or the JQuery framework and its plugins. The code developed is also open source and the server deployment could be achieved easily within minutes. Two mirrors currently exist, one in Madrid at the Centro de Biology Molecular Severo Ochoa and the other at the ChEMBL group at the European Bioinformatics Institute (EBI) in Cambridge. The latter is also linked to the Protein Data Bank (PDB) web interface. This allows any user to obtain direct information regarding a complex if it is available. In the PDB web interface there is also a

link for each complex to represent the CBS corresponding to a complex in the AtlasCBS server.

As a web interface, the AtlasCBS has some technical limitations. In the first place, the speed and the number of compounds that can be loaded simultaneously strongly depend on the browser used by the user. As the application heavily relies on AJAX, JavaScript runtime engines become crucial. Secondly, data upload is also limited to one thousand compounds, which should be enough for most of the possible server uses. The AtlasCBS, despite its limitations, is a powerful tool that has been proved to provide excellent result analyzing past drug discovery projects and identifying optimal design paths. Future developments may solve the pending issues.

Finally, in the last methodological piece of work, a *de novo* fragment-based design protocol has been implemented and tested using some successful cases available from the literature. We tested that the ligand efficiency framework is able to drive the optimization process. However, the nature of the protocol is still dependent on the user's choices and decisions, which is not the ideal scenario.

Unlike other parameters such as binding energy, ligand efficiency indices have not been analyzed for the optimal range of values in the optimization process, e.g. larger SEI (polarity) values may not be better than smaller ones, but different regions of the scale could be necessary for different targets. More work is needed in this regard to provide a comprehensive view of the scales and the range of values that could optimize the paths.

Experimental validation is limited and was kindly provided by other groups. Two projects have provided the highest degree of evidence supporting the methods utility. The first one is the FtsZ VS project. In this case, docking, pharmacophore search and very detailed energy analysis were combined to produce a short list of 5 possible candidates that could bind to the intended target. Four out of five compounds were determined to bind to the FtsZ protein.

The second one was the determination of the binding mode of substrates of a tannase enzyme from *Lactobacillus plantarum*, thanks to a collaboration with Dr. José Miguel Mancheño's group at the Instituto de Química Física Rocasolano (CSIC, Madrid). The prediction of the binding mode was confirmed later when the X-ray structures of several complexes were determined by Qianming Chen et al. (Ren, Wu et al. 2013)

Other modeling projects have employed the tools described in this thesis with success, but the results from these have not been experimentally tested or validated yet.

6. CONCLUSIONES

1. Se ha desarrollado una interfaz gráfica, métodos basados en ligandos y se ha actualizado la plataforma de cribado actual VSDMIP.

2. Se han establecido diferentes protocolos para acortar los tiempos de cribado virtual y mejorar diversidad de los resultados.

3. El desarrollo del programa de *docking* CRDOCK ha supuesto una mejora en la precisión y velocidad de los cribados virtuales.

4. Se ha desarrollado la herramienta AtlasCBS que permite el análisis del espacio químico-biológico para optimización de compuestos.

5. Se ha desarrollado una herramienta para la exploración de efectos adversos y perfiles de polifarmacología.

.

7. REFERENCIAS

Abad-Zapatero, C. (2007). "Ligand efficiency indices for effective drug discovery."
Abad-Zapatero, C. and D. Blasi (2011). "Ligand Efficiency Indices (LEIs): more than a simple efficiency yardstick." Molecular Informatics **30**(2-3): 122-132.
Abad-Zapatero, C., O. Perisic, et al. (2010). "Ligand efficiency indices for an effective mapping of chemico-biological space: the concept of an atlas-like representation." Drug discovery today **15**(19): 804-811.
Adams, G. P. and L. M. Weiner (2005). "Monoclonal antibody therapy of cancer." Nature biotechnology **23**(9): 1147-1157.
Alder, B. J. and T. Wainwright (1959). "Studies in molecular dynamics. I. General method." The Journal of Chemical Physics **31**: 459.
Allen, F. H. (2002). "The Cambridge Structural Database: a quarter of a million crystal structures and rising." Acta Crystallographica Section B: Structural Science **58**(3): 380-388.
Andersen, H. C. (1983). "RATTLE: A "Velocity" version of the SHAKE algorithm for molecular dynamics calculations." Journal of Computational Physics **52**(1): 24-34.
Armstrong, M. S., G. M. Morris, et al. (2010). "ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics." Journal of computer-aided molecular design **24**(9): 789-801.
Arrell, D. and A. Terzic (2010). "Network systems biology for drug discovery." Clinical Pharmacology & Therapeutics **88**(1): 120-125.
Baber, J. C., D. C. Thompson, et al. (2009). "GARD: a generally applicable replacement for RMSD." Journal of chemical information and modeling **49**(8): 1889-1900.
Ballester, P. J. and W. G. Richards (2007). "Ultrafast shape recognition to search compound databases for similar molecular shapes." Journal of computational chemistry **28**(10): 1711-1723.
Barril, X. and F. J. Luque (2012). "Molecular simulation methods in drug discovery: a prospective outlook." Journal of computer-aided molecular design **26**(1): 81-86.
Baum, B., L. Muley, et al. (2009). "Think twice: understanding the high potency of bis (phenyl) methane inhibitors of thrombin." Journal of molecular biology **391**(3): 552-564.
Bembenek, S. D., B. A. Tounge, et al. (2009). "Ligand efficiency and fragment-based drug discovery." Drug discovery today **14**(5): 278-283.
Bemis, G. W. and M. A. Murcko (1996). "The properties of known drugs. 1. Molecular frameworks." Journal of medicinal chemistry **39**(15): 2887-2893.
Berger, S. I. and R. Iyengar (2009). "Network analyses in systems pharmacology." Bioinformatics **25**(19): 2466-2472.
Berman, H. M., J. Westbrook, et al. (2000). "The protein data bank." Nucleic acids research **28**(1): 235-242.
Berthold, M. R., N. Cebron, et al. (2008). KNIME: The Konstanz information miner, Springer.
Blasi, D., G. Arsequell, et al. (2011). "Retrospective Mapping of SAR Data for TTR Protein in Chemico-Biological Space Using Ligand Efficiency Indices as a Guide to Drug Discovery Strategies." Molecular Informatics **30**(2-3): 161-167.
Boehm, M., T.-Y. Wu, et al. (2008). "Similarity searching and scaffold hopping in synthetically accessible combinatorial chemistry spaces." Journal of medicinal chemistry **51**(8): 2468-2480.
Böhm, H.-J. (1992). "The computer program LUDI: a new method for the de novo design of enzyme inhibitors." Journal of computer-aided molecular design **6**(1): 61-78.
Boobbyer, D. N., P. J. Goodford, et al. (1989). "New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure." Journal of medicinal chemistry **32**(5): 1083-1094.

Born, M. and R. Oppenheimer (1927). "Zur quantentheorie der molekeln." Annalen der Physik **389**(20): 457-484.

Brenk, R., L. Naerum, et al. (2003). "Virtual screening for submicromolar leads of tRNA-guanine transglycosylase based on a new unexpected binding mode detected by crystal structure analysis." Journal of medicinal chemistry **46**(7): 1133-1143.

Cabrera, Á. C., R. Gil-Redondo, et al. (2011). "VSDMIP 1.5: an automated structure-and ligand-based virtual screening platform with a PyMOL graphical user interface." Journal of computer-aided molecular design **25**(9): 813-824.

Cabrera, Á. C., J. Klett, et al. (2012). "CRDOCK: An Ultrafast Multipurpose Protein-Ligand Docking Tool." Journal of chemical information and modeling **52**(8): 2300-2309.

Congreve, M., R. Carr, et al. (2003). "A "rule of three" for fragment-based lead discovery?" Drug discovery today **8**(19): 876-877.

Congreve, M., G. Chessari, et al. (2008). "Recent developments in fragment-based drug discovery." Journal of medicinal chemistry **51**(13): 3661-3680.

Connelly, P. R., R. A. Aldape, et al. (1994). "Enthalpy of hydrogen bond formation in a protein-ligand binding reaction." Proceedings of the National Academy of Sciences **91**(5): 1964-1968.

Cornell, W. D., P. Cieplak, et al. (1995). "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules." Journal of the American Chemical Society **117**(19): 5179-5197.

Cross, J. B., D. C. Thompson, et al. (2009). "Comparison of several molecular docking programs: pose prediction and virtual screening accuracy." Journal of chemical information and modeling **49**(6): 1455-1474.

ChemAxon. (2013). "ChemAxon web page." from http://www.chemaxon.com/.

Chen, H.-M., B.-F. Liu, et al. (2007). "SODOCK: Swarm optimization for highly flexible protein-ligand docking." Journal of computational chemistry **28**(2): 612-623.

Darden, T., D. York, et al. (1993). "Particle mesh Ewald: An N log (N) method for Ewald sums in large systems." The Journal of Chemical Physics **98**: 10089.

Dolinsky, T. J., J. E. Nielsen, et al. (2004). "PDB2PQR: an automated pipeline for the setup of Poissonâ€“Boltzmann electrostatics calculations." Nucleic acids research **32**(suppl 2): W665-W667.

Donati, C. and R. Rappuoli (2013). "Reverse vaccinology in the 21st century: improvements over the original design." Annals of the New York Academy of Sciences **1285**: 115-132.

Eckert, H. and J. r. Bajorath (2007). "Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches." Drug discovery today **12**(5): 225-233.

Eldridge, M. D., C. W. Murray, et al. (1997). "Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes." Journal of computer-aided molecular design **11**(5): 425-445.

Fan, H., D. Schneidman-Duhovny, et al. (2011). "Statistical potential for modeling and ranking of proteinâ€“ligand interactions." Journal of chemical information and modeling **51**(12): 3078-3092.

Fechner, U. and G. Schneider (2006). "Flux (1): a virtual synthesis scheme for fragment-based de novo design." Journal of chemical information and modeling **46**(2): 699-707.

Fischer, E. (1894). "Einfluss der Configuration auf die Wirkung der Enzyme." Berichte der deutschen chemischen Gesellschaft **27**(3): 2985-2993.

Fletcher, R. (1980). Practical methods of optimization, John Wiley & Sons.

Friesner, R. A., J. L. Banks, et al. (2004). "Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy." Journal of medicinal chemistry **47**(7): 1739-1749.

Gaulton, A., L. J. Bellis, et al. (2012). "ChEMBL: a large-scale bioactivity database for drug discovery." Nucleic acids research **40**(D1): D1100-D1107.

Goodford, P. J. (1985). "A computational procedure for determining energetically favorable binding sites on biologically important macromolecules." Journal of medicinal chemistry **28**(7): 849-857.

Götz, A. W., M. J. Williamson, et al. (2012). "Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized born." Journal of Chemical Theory and Computation **8**(5): 1542.

Grant, J. A., M. Gallardo, et al. (1996). "A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape." Journal of computational chemistry **17**(14): 1653-1666.

Halgren, T. A. (1996). "Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94." Journal of computational chemistry **17**(5â€6): 490-519.

Hann, M. M. and G. M. Keserü (2012). "Finding the sweet spot: the role of nature and nurture in medicinal chemistry." Nature reviews Drug discovery **11**(5): 355-365.

Head, J. D. and M. C. Zerner (1985). "A Broyden-Fletcher-Goldfarb-Shanno optimization procedure for molecular geometries." Chemical physics letters **122**(3): 264-270.

Hess, B. (2008). "P-LINCS: A parallel linear constraint solver for molecular simulation." Journal of Chemical Theory and Computation **4**(1): 116-122.

Hestenes, M. R. and E. Stiefel (1952). "Methods of conjugate gradients for solving linear systems." Journal of Research of the National Bureau of Standards **49**(6): 409-436.

Hoshino, R., Y. Chatani, et al. (1999). "Constitutive activation of the 41-/43-kDa mitogen-activated protein kinase signaling pathway in human tumors." Oncogene **18**(3): 813-822.

Hou, T., J. Wang, et al. (2010). "Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations." Journal of chemical information and modeling **51**(1): 69-82.

Huang, Q., L.-L. Li, et al. (2010). "PhDD: A new pharmacophore-based *de novo* design method of drug-like molecules combined with assessment of synthetic accessibility." Journal of Molecular Graphics and Modelling **28**(8): 775-787.

Irwin, J. J. (2008). "Community benchmarks for virtual screening." Journal of computer-aided molecular design **22**(3-4): 193-199.

Irwin, J. J. and B. K. Shoichet (2005). "ZINC-a free database of commercially available compounds for virtual screening." Journal of chemical information and modeling **45**(1): 177-182.

Jhoti, H., G. Williams, et al. (2013). "The 'rule of three' for fragment-based drug discovery: where are we now?" Nature reviews Drug discovery **12**(8): 644-644.

Kishnani, P. S., R. D. Steiner, et al. (2006). "Pompe disease diagnosis and management guideline." Genetics in Medicine **8**(5): 267-288.

Klett, J., A. Nuñez-Salgado, et al. (2012). "MM-ISMSA: An Ultrafast and Accurate Scoring Function for Protein-Protein Docking." Journal of Chemical Theory and Computation **8**(9): 3395-3408.

Korb, O., T. Stutzle, et al. (2009). "Empirical scoring functions for advanced protein-ligand docking with PLANTS." Journal of chemical information and modeling **49**(1): 84-96.

Koshland Jr, D. (1958). "Application of a theory of enzyme specificity to protein synthesis." Proceedings of the National Academy of Sciences of the United States of America **44**(2): 98.

Kuntz, I. D., J. M. Blaney, et al. (1982). "A geometric approach to macromolecule-ligand interactions." Journal of molecular biology **161**(2): 269-288.

Kutchukian, P. S., D. Lou, et al. (2009). "FOG: Fragment optimized growth algorithm for the de novo generation of molecules occupying druglike chemical space." Journal of chemical information and modeling **49**(7): 1630-1642.

Lewell, X. Q., D. B. Judd, et al. (1998). "RECAP retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry." Journal of chemical information and computer sciences **38**(3): 511-522.

Lipinski, C. A., F. Lombardo, et al. (1997). "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings." Advanced drug delivery reviews **23**(1): 3-25.

Liu, M.-M., L. Zhou, et al. (2012). "Discovery of flavonoid derivatives as anti-HCV agents via pharmacophore search combining molecular docking strategy." European journal of medicinal chemistry **52**: 33-43.

Liu, T., Y. Lin, et al. (2007). "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities." Nucleic acids research **35**(suppl 1): D198-D201.

Liu, Z., G. Wang, et al. (2008). "Geometrical Preferences of the Hydrogen Bonds on Proteinâˆ' Ligand Binding Interface Derived from Statistical Surveys and Quantum Mechanics Calculations." Journal of Chemical Theory and Computation **4**(11): 1959-1973.

Lounkine, E., M. J. Keiser, et al. (2012). "Large-scale prediction and testing of drug activity on side-effect targets." Nature **486**(7403): 361-367.

Ma, B., S. Kumar, et al. (1999). "Folding funnels and binding mechanisms." Protein Engineering **12**(9): 713-720.

Massova, I. and P. A. Kollman (2000). "Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding." Perspectives in drug discovery and design **18**(1): 113-135.

Mayo, S. L., B. D. Olafson, et al. (1990). "DREIDING: a generic force field for molecular simulations." Journal of Physical Chemistry **94**(26): 8897-8909.

McGann, M. (2012). "FRED and HYBRID docking performance on standardized datasets." Journal of computer-aided molecular design **26**(8): 897-906.

Morgan, H. (1965). "The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service." Journal of Chemical Documentation **5**(2): 107-113.

Morreale, A., R. Gil-Redondo, et al. (2007). "A new implicit solvent model for protein-ligand docking." PROTEINS: Structure, Function, and Bioinformatics **67**(3): 606-616.

Morris, G. M., R. Huey, et al. (2009). "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility." Journal of computational chemistry **30**(16): 2785-2791.

Munos, B. (2009). "Lessons from 60 years of pharmaceutical innovation." Nature reviews Drug discovery **8**(12): 959-968.

Murray, C. W. and D. C. Rees (2009). "The rise of fragment-based drug discovery." Nature chemistry **1**(3): 187-192.

Nagle, T., C. Berg, et al. (2003). "The further evolution of biotech." Nature reviews Drug discovery **2**(1): 75-79.

Nelder, J. A. and R. Mead (1965). "A simplex method for function minimization." The computer journal **7**(4): 308-313.

Neves, M. A., M. Totrov, et al. (2012). "Docking and scoring with ICM: the benchmarking results and strategies for improvement." Journal of computer-aided molecular design **26**(6): 675-686.

O'Boyle, N. M., M. Banck, et al. "Open Babel: An open chemical toolbox." Journal of cheminformatics **3**(1): 1-14.

O'Boyle, N. M., C. Morley, et al. (2008). "Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit." Chem Cent J **2**(5).

Oprea, T. I., J. E. Bauman, et al. (2012). "Drug repurposing from an academic perspective." Drug Discovery Today: Therapeutic Strategies **8**(3): 61-69.

Pammolli, F., L. Magazzini, et al. (2011). "The productivity crisis in pharmaceutical R&D." Nature reviews Drug discovery **10**(6): 428-438.

Pande, V. S., I. Baker, et al. (2003). "Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing." Biopolymers **68**(1): 91-109.

Pérez, C. and A. R. Ortiz (2001). "Evaluation of docking functions for protein-ligand docking." Journal of medicinal chemistry **44**(23): 3768-3785.

Petrova, S. S. and A. D. Solov'ev (1997). "The origin of the method of steepest descent." Historia Mathematica **24**(4): 361-375.

Pilot, P. "Version 8.5. Accelrys." Inc.: San Diego, CA.

Politzer, P., J. S. Murray, et al. (2010). "Halogen bonding: an electrostatically-driven highly directional noncovalent interaction." Physical Chemistry Chemical Physics **12**(28): 7748-7757.

Proschak, E., H. Zettl, et al. (2009). "From molecular shape to potent bioactive agents I: bioisosteric replacement of molecular fragments." ChemMedChem **4**(1): 41-44.

Ren, B., M. Wu, et al. (2013). "Crystal structure of tannase from *Lactobacillus plantarum*." Journal of molecular biology **425**(15): 2737–2751.

Reulecke, I., G. Lange, et al. (2008). "Towards an integrated description of hydrogen bonding and dehydration: decreasing false positives in virtual screening with the HYDE scoring function." ChemMedChem **3**(6): 885-897.

Reynolds, C. H., B. A. Tounge, et al. (2008). "Ligand binding efficiency: trends, physical basis, and implications." Journal of medicinal chemistry **51**(8): 2432-2438.

Rogers, D. and M. Hahn (2010). "Extended-connectivity fingerprints." Journal of chemical information and modeling **50**(5): 742-754.

Rogers, D. J. and T. T. Tanimoto (1960). "A computer program for classifying plants." Science **132**(3434): 1115-1118.

Ruppert, J., W. Welch, et al. (1997). "Automatic identification and representation of protein binding sites for molecular docking." Protein Science **6**(3): 524-533.

Rush, T. S., J. A. Grant, et al. (2005). "A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction." Journal of medicinal chemistry **48**(5): 1489-1495.

Sándor, M., R. Kiss, et al. (2010). "Virtual fragment docking by Glide: A validation study on 190 protein-fragment complexes." Journal of chemical information and modeling **50**(6): 1165-1172.

Saxty, G., S. J. Woodhead, et al. (2007). "Identification of inhibitors of protein kinase B using fragment-based lead discovery." Journal of medicinal chemistry **50**(10): 2293-2296.

Scannell, J. W., A. Blanckley, et al. (2012). "Diagnosing the decline in pharmaceutical R&D efficiency." Nature reviews Drug discovery **11**(3): 191-200.

Schneider, G., W. Neidhart, et al. (1999). ""Scaffold Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening." Angewandte Chemie International Edition **38**(19): 2894-2896.

Schneider, N., G. Lange, et al. (2013). "A consistent description of HYdrogen bond and DEhydration energies in protein-ligand complexes: methods behind the HYDE scoring function." Journal of computer-aided molecular design **27**(1): 15-29.

Schrodinger, L. "PyMOL molecular graphics system, version 1.5. 0.4." See http://pymol.org.
Senger, S. (2009). "Using Tversky similarity searches for core hopping: finding the needles in the haystack." Journal of chemical information and modeling **49**(6): 1514-1524.
Shaw, D. E., M. M. Deneroff, et al. (2007). Anton, a special-purpose machine for molecular dynamics simulation. ACM SIGARCH Computer Architecture News, ACM.
Sheng, C. and W. Zhang (2012). "Fragment Informatics and Computational Fragments□Based Drug Design: An Overview and Update." Medicinal Research Reviews.
Sherman, W., T. Day, et al. (2006). "Novel procedure for modeling ligand/receptor induced fit effects." Journal of medicinal chemistry **49**(2): 534-553.
Strebhardt, K. and A. Ullrich (2008). "Paul Ehrlich's magic bullet concept: 100 years of progress." Nature Reviews Cancer **8**(6): 473-480.
Totrov, M. and R. Abagyan (2008). "Flexible ligand docking to multiple receptor conformations: a practical alternative." Current opinion in structural biology **18**(2): 178-184.
Treiber, D. K. and N. P. Shah (2013). "Ins and Outs of Kinase DFG Motifs." Chemistry & Biology **20**(6): 745-746.
Truchon, J.-F. and C. I. Bayly (2007). "Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem." Journal of chemical information and modeling **47**(2): 488-508.
Tsai, C. J., S. Kumar, et al. (1999). "Folding funnels, binding funnels, and protein function." Protein Science **8**(6): 1181-1190.
Vanderbilt, D. and S. G. Louie (1984). "A Monte Carlo simulated annealing approach to optimization over continuous variables." Journal of Computational Physics **56**(2): 259-271.
Varin, T., R. Bureau, et al. (2009). "Clustering files of chemical structures using the Székely-Rizzo generalization of Ward's method." Journal of Molecular Graphics and Modelling **28**(2): 187-195.
Verdonk, M. L., J. C. Cole, et al. (2003). "Improved protein-ligand docking using GOLD." PROTEINS: Structure, Function, and Bioinformatics **52**(4): 609-623.
Wang, B. and K. M. Merz Jr (2010). "Importance of loop dynamics in the neocarzinostatin chromophore binding and release mechanisms." Physical Chemistry Chemical Physics **12**(14): 3443-3449.
Wang, J., R. M. Wolf, et al. (2004). "Development and testing of a general amber force field." Journal of computational chemistry **25**(9): 1157-1174.
Wang, P. C. and T. E. Shoup (2011). "Parameter sensitivity study of the Nelder-Mead simplex method." Advances in Engineering Software **42**(7): 529-533.
Wang, R., X. Fang, et al. (2004). "The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures." Journal of medicinal chemistry **47**(12): 2977-2980.
Warren, G. L., C. W. Andrews, et al. (2006). "A critical assessment of docking programs and scoring functions." Journal of medicinal chemistry **49**(20): 5912-5931.
Yuan, Y., J. Pei, et al. (2011). "LigBuilder 2: a practical de novo drug design approach." Journal of chemical information and modeling **51**(5): 1083-1091.
Zhang, Y. and J. Skolnick (2004). "Scoring function for automated assessment of protein structure template quality." PROTEINS: Structure, Function, and Bioinformatics **57**(4): 702-710.
Zhao, S. and R. Iyengar (2012). "Systems pharmacology: network analysis to identify multiscale mechanisms of drug action." Annual review of pharmacology and toxicology **52**: 505.